

What happened to all that prosperity?

Fears that reform in the former Soviet Union will be impeded by ignorance of market economics should be matched by the suspicion that the policies of experienced free market governments are undermined by ignorance of arithmetic.

ECONOMISTS, practitioners of what is called the 'dismal science', are having a rotten time, as usual. Just as architects sometimes rashly claim to be able to engineer the physical environment so as to make people happy, so economists are expected to engineer the monetary environment so as to make the rest of us prosperous.

But even when they are able to engineer a period of economic growth, such as that which lasted from about 1982 to just after the stock market crash in October 1987, sustained growth seems to elude them. Now, with many of the world's industrialized nations seeking escape from a recession that seems unexpectedly obdurate, economists as a group seem to have no clear remedy, to the intense irritation of politicians, especially those seeking re-election this year. The reminder that there is little in the way of success to report after decades of struggling with the problems of the developing countries may seem unfair, but is not irrelevant.

The first thing to say is that all these charges are unfair. So much can be told from what is happening in the United States and Britain, whose governments seek re-election in November and before July respectively, and would prefer to do so on a rising tide of prosperity. Each is seeking ways of moderating the policies economists would recommend by stratagems calculated to make them more electable. The people to blame are the politicians, not the economists.

Thus President George Bush is widely expected to make the centrepiece of his State of the Union address later this month a "growth package" designed to "kick-start" the US economy. Although the Federal Reserve's discount rate (which fell to 4.5 per cent last month) is the lowest for a quarter of a century, Bush can only throw more money at the recession, increasing the chances that the federal deficit will rise above the horrendous \$500,000 million expected for this year. And whatever the outcome of the tough talking on trade advertised for Tokyo this week, it seems improbable that there can be any substantial physical growth of US exports to Japan (and thus an increase of domestic manufacturing) before the election.

In Britain (as usually) the shoe is on the other foot. Now that Britain formally belongs to the European Exchange Rate Mechanism (ERM), the classical escape from financial crisis (letting sterling depreciate against other currencies) is denied. The immediate embarrassment is last month's decision of the German central bank to increase interest rates by a full one per cent. If there had

been no election in prospect, British interest rates would probably also have been increased (as they have been in France and elsewhere in Europe), but that would make quick recovery from the recession even less likely. So the government seems to have firmly crossed its fingers, hoping for a little luck. Its promise that it will not devalue will no doubt be requited by the run-down of the reserves, managed by the Bank of England.

Long before the election comes, the government may be wishing that it had congratulated itself less fervently on its hollow victory at Maastricht last month, when it won the concession from its partner governments that Britain might opt out of the end-point of the ERM — a single currency for the whole of Europe, and a single central bank. The truth is that if this apparatus were now in place, Britain's present problems would be more manageable than they are. An effective European central bank might, for example, have persuaded the German government to deal with the inflationary pressures of reunification by increasing taxes, not interest rates. Otherwise, the drain on Britain's reserves now expected in the defence of sterling would have been met by flows of funds for other purposes from Europe centrally.

Short-term interests prevail

The lesson seems to be that, in the interaction between politics and economics, the short-term interests of politicians tempt them to select from the remedies suggested by economists only the more palatable elements. Bush evidently finds the federal deficit an irksome constraint, so much so that he seems willing to let Japan think of him as a high-powered motor-car salesman, but (like his predecessor) cannot face the unpalatable consequence that the deficit could be made to vanish only by impoverishing each resident of the United States by something like \$2,500 a year. Meanwhile, the present metastable state of affairs is maintained only by the substantial external deficit, which puts money into the pockets of people elsewhere which they promptly transfer back to the United States, either by making investments of lending to the government (by the purchase of US Treasury bonds). So much is simple but forgotten arithmetic.

Also forgotten is the arithmetic of the recession. A year ago, the world's financial system seemed on the edge of collapse, with major banks in the United States reporting



huge losses on loans to developing countries, with the US Congress making arrangements to borrow \$500,000 million to rescue the savings and loan industry and with the prospect that the rate of failure among smaller banks would continue apace. The rot has been stopped, but at considerable cost. Banks have become cautious in their lending policies (and have earned brickbats from the White House for what previously would have been regarded as admirable prudence), while borrowers have similarly been made shy of debt by the abounding tales of how tough are banks when bent on the repossession of people's houses and motor-cars.

Hidden in this is the evaporation of much wealth. Banks that have lent funds to property developers for the construction of empty office blocks may comfort themselves that the buildings on which they have a lien may one day become the 'performing assets' they were intended to be, but that is by no means certain. In any case, in the interval, the funds used in their construction are sterilized, and will be out of economic action for many years. The assets of the many companies that have spectacularly gone bust in recent months have often similarly been taken out of action, as will be much past investment in the 21 manufacturing plants General Motors intends to close in the next few years.

Hungry for capital

The pity — and another reason why the recession appears so resistant to treatment — is that all this has happened when the world as a whole is hungry for capital. Estimates vary, but putting the ex-Soviet republics on their feet may require the investment of some \$200,000 million a year for 20 years or so, the eastern part of Germany is absorbing investment at about a tenth of that rate (but will probably be self-sustaining sooner) — and there remains the whole of the developing world, now almost forgotten again. It is no wonder that there is what Western politicians call a 'credit crunch'. On present form, it could last for years, even decades. The obvious danger, for countries such as the United States and Britain, is that their capacity to compete with investment funds in what has become a global market will be permanently diminished. So much is simple arithmetic.

Is there an escape? The best hope, perhaps the only immediate hope, is that the revision of the General Agreement on Tariffs and Trade (GATT), under negotiation for the past five years, will eventually come to something. In circumstances as serious as the present, it is an assault on common sense that the only device in sight for increasing the volume of the world's economic activity should be stalled because of the attachment of Europe, the United States and Japan (in descending order of culpability) to systems of agricultural subsidy that are ruinous for their own taxpayers and harmful to everybody else. That is also a matter of arithmetic, too often disguised as national strategic interest and too often forgotten because farmers appear exempt from the old rule of one man (or woman), one vote. □

Paying for research

The real question is whether the US government should bear the full costs of academic research.

THE US government's heart-searching about the overhead costs of university research continues. Unavoidably, much attention has been focused on the details of accounting systems. Should Stanford University, the hapless fall-guy in the past year's scandal, have organized its bookkeeping so that certain indirect costs — those for its president's house, for example — were automatically excluded from the pool of costs charged to research budgets? Obviously, the answer is 'Yes'.

A more important question, raised by the government's current attempt to renege on special agreements with Stanford and other research universities (see p. 97), is whether federal officials are being fair when they seek to recover costs already legally allocated to the indirect-cost pool. The answer here has two parts. First, it is not fair for the government to change the rules retrospectively and, second, it is damaging to the interests of research in general that university research budgets once agreed should then be put in jeopardy.

But the real issue arising from the indirect-costs scandal is the extent to which the US government should fund university research. Is the objective full reimbursement? Or just part of the whole cost? There is no doubt but that universities, even when large and well endowed, lose money on research grants. The government, contrary to general belief, has not been providing full reimbursement of indirect costs even under the old rules. It is therefore understandable that universities such as Stanford and the Massachusetts Institute of Technology (MIT) have aggressively set about charging everything they legally might to the indirect-costs pool.

So what degree of reimbursement is proper? One argument for parsimony is that it is perfectly appropriate for universities to pay a share of the costs of academic endeavours that enrich their intellectual life and also bring a good deal of prestige. But where should the unrequited costs of research then be charged? Against undergraduate teaching, thus increasing tuition fees at a time when the recruitment of young people to science studies is generally acknowledged to be an urgent need? Or against the generosity of alumni and other benefactors, and at the cost of the general development of an institution? And what, then, would happen to predominantly research institutions, of which the Rockefeller University is a prime example?

The truth is that research is in the national interest. That is why the full cost of research is the proper responsibility of the federal government. That this question has not been resolved, and is not even being discussed openly, is one reason why the system for research support in the United States is under needless stress. The more tantalizing issue of which items are charged to the cost pool may excite the interest of bystanders and the indignation of congressmen, but is by comparison unimportant. □

US government targets indirect cost agreements

- Auditors claim exceptions were error
- Stanford, MIT cry foul, promise court battle

Washington

GOVERNMENT auditors are threatening to disallow dozens of special research cost agreements with several US universities, a move that could cost the institutions millions of dollars in lost overhead cost reimbursements.

Claiming that government officials had mistakenly entered into special agreements with Stanford University, the Massachusetts Institute of Technology (MIT), the California Institute of Technology, the University of Hawaii and others, auditors at the Defense Contract Audit Agency (DCAA) are trying to have the agreements invalidated. Known as Memoranda of Understanding (MoU), the agreements exempted a small number of universities from normal accounting procedures, to allow them to recover special overhead costs unique to their institutions.

A report completed last week by DCAA, which is responsible for negotiating and auditing overhead reimbursement at about 40 universities, recommended that MIT withdraw about \$22 million of its projected \$131 million reimbursement request. Some \$3 million of that is due to a MoU, which expired in 1990, covering the library. James Culliton, MIT's vice president for finance, says that he had expected that the agreement would be continued while university officials conduct a new study of real library costs; instead DCAA has recommended cutting the library reimbursement rate in half.

Another dispute centres around the cost of the Lincoln Laboratory, a research institution that has been run by MIT since the Second World War. DCAA intends to separate accounting for the two institutions, which would halt the traditional practice of charging MIT administrative overhead to Lincoln Laboratory research budgets. The move could lose the university about \$8 million this year. "We'll take that straight to the court of appeals," warns Culliton. "This is a radical change to an old relationship."

Stanford University is predicting that a similar review at that institution could have even more serious implications. University officials say that a DCAA audit has questioned approximately \$300 million in research costs charged in the 1980s. "To arrive at such a high number," said chief financial officer Peter Van Etten in a statement, "DCAA must have disregarded binding, written contracts between Stanford and the government." Long term MoU

had allowed Stanford to bypass certain record-keeping regulations. But now that those agreements have been questioned, Stanford is afraid that auditors will allow no reimbursement at all, on the Catch-22 argument that the university had not kept proper records.

"The heart of Stanford's dispute with DCAA is not yachts and flowers; Stanford voluntarily withdrew such costs many months ago," Van Etten said. "Rather, the dispute is about the fair, actual costs of supporting research and the government's contractual agreements to pay for those costs." Since the indirect cost scandal broke in 1990, Stanford has repaid \$2.3 million in questioned research costs.

Unlike the MIT audit, in which MoU were only disallowed for current and future costs, DCAA appears to be retroactively questioning all of Stanford's agreements for the past decade. Some of the agreements at Stanford were intended to tailor the standard indirect cost formulas to the actual conditions at the university. Rather than assume the same utility rate per square foot in all campus buildings, for example, a special MoU accounts for the higher utility costs in research laboratories. Last April, the Navy, which oversees the indirect cost agreements that are now under review, cancelled all the Stanford MoU, which effectively lowered the university's recoverable indirect cost rate from 70 to 55.5 per cent of direct research costs. University lawyers filed suit over that move and threaten to do the same if the government attempts to disallow agreements made in the 1980s.

Congressional investigators on the oversight and investigations subcommittee of Representative John Dingell (Democrat, Michigan), who has led the congressional attack against indirect cost abuses, say that the issue stems from fundamental lapses in federal accounting oversight in the past decade. "Government bureaucrats didn't follow their own regulations when entering into these agreements. Nobody was protecting the government's interests," says Bruce Chafin, a Dingell staff member. The Navy has disciplined the accountants who negotiated the original contracts, he says.

Other universities that entered into MoU with federal agencies are still under audit. Caltech spokesman Hall Daily says DCAA is yet to question any of that university's dozen MoU, but that the agency has increased the number of auditors examining Caltech accounts and should

be finishing a report soon.

DCAA is circulating its Stanford and MIT reports to federal science agencies, seeking comment before making a final decision in the cases by 31 January.

Dingell plans to hold a hearing on 30 January, to hear from the various agencies that have jurisdiction over indirect cost reimbursement. Auditors from the DCAA, the Navy's Office of Naval Research, the department of Health and Human Services (which negotiates indirect cost rates with the majority of US universities) and investigators from the congressional General Accounting Office are expected to testify. **Christopher Anderson**

BRITISH SCIENCE

Appeal across water

London

THE pressure group British Scientists Abroad (BSA) has taken to the worlds' electronic bulletin boards to urge British scientists to mount a letter-writing campaign meant to raise the profile of science policy in the run up to the general election.

BSA, which asserts that present government policies are contributing to a serious 'brain drain' of British scientists, most of whom end up in the United States, wants British researchers — whether at home or abroad — to write to their local member of parliament expressing concern about the state of British science and engineering. The appeal has been posted on a number of the leading global electronic bulletin boards used by research scientists.

Michael Duff, a British physicist at Texas A&M University, hopes that researchers will compose their own letters, but BSA has suggested a draft, addressed to Alan Howarth, minister for science and higher education, which argues that government policies have for many years neglected science, citing a survey which found that only 6.5 per cent of British university teachers believe that scientific research is an attractive career for British graduates.

As a result, the impact of British scientists' publications is declining, the letter claims. BSA wants a substantial increase in government spending on science, but can expect a frosty response to this request from the Conservative government, which says that the state of British science should be measured by its output — always a difficult quantity to measure — rather than by the proportion of British wealth spent on research. The Labour opposition has also made no specific promises about the amount of money that will be made available for science, should it win the election, saying that this will depend on the degree of growth in the British economy.

Peter Aldhous

East Europe's reactors in trouble

AN international study of the basic design faults of more than 20 ageing Soviet-built pressurized water reactors in Eastern Europe has given rise to great concern about safety.

The state of the nuclear power industry of that vulnerable and densely populated region has now been reviewed in a scientific programme mounted by the International Atomic Energy Agency (IAEA). The study identified more than 1,000 specific problems arising from weaknesses that could lead to disaster.

The study was part of a larger survey of 62 largely obsolete Soviet-designed nuclear power plants, most of them in Europe. "It is absolutely necessary," a specialist IAEA spokesman emphasizes in diplomatic language barely masking the alarm caused by the findings, "that the safety concerns should be addressed with reasonable economic means and within an acceptable time frame." One result of the expert report is that Western nuclear energy companies, much of whose business may have been lost after the Chernobyl disaster in the former Soviet Union, may expect a steady flow of orders from Eastern Europe for grafting new technology onto old reactors in order to prolong their working lives in relative safety.

The outcome of the review was discussed at the end of last year at a private meeting at Vienna attended by specialists from the IAEA member states. They concluded that nuclear power would

continue to play a dominant role in Eastern Europe because of the dependence of its industries on imported fossil energy as well as degradation of the environment.

The study described some strengths and the many drawbacks of the WWER 440/230 design concept. (The Chernobyl reactor that malfunctioned was a water-cooled graphite-moderated design known as RKB1000). It identified basic weaknesses caused by inadequacies of instrumentation and control as well as electric power supply. The review also complained that the WWER designers had carried out analyses of only a narrow spectrum of potential accidents.

The embrittlement of pressure vessels has given cause for particular concern. So have the quality and practice of operation and maintenance activities, protection against external hazards and fires. The investigators fear that the problems of developing the proper attitude to safety and quality essential for running a potentially dangerous industry may well prove difficult — although not as expensive as the essential technical improvements that must be undertaken.

One logical conclusion of the study is that acute and potentially disastrous weaknesses at other reactors are probably still waiting to be identified. (A separate, confidential report drawn up by the US Central Intelligence Agency considers that at least four reactors in the former Soviet Union are functioning in safety conditions

even worse than those prevailing at Chernobyl before the accident there.)

But the experts also agreed that recent political changes and the recent combined work of the IAEA, the European Communities, the Organization for Economic Cooperation and Development, the Nuclear Energy Agency and the World Association of Nuclear Operators, as well as a large number of bilateral efforts, had created the conditions for large-scale cooperation essential for the task at hand.

A basic design fault of the WWER model is held responsible for a near meltdown in 1975 at a four-unit 1,700 MW reactor complex at the Baltic Sea town of Greifswald, in the former East Germany. The incident became known only after the publication of East German state secrets. It seems that a disaster perhaps comparable to that at Chernobyl was narrowly averted when fire cut off power to 11 of the 12 cooling pumps serving one of the reactors at the plant.

Spares from the condemned Greifswald plant are destined for Kozloduy in northern Bulgaria, two of whose still-functioning reactors are reckoned to be 1,000 times more likely to cause a devastating disaster than their modern equivalents in the West. The Bulgarian plant operators understand the risk, but told the Vienna meeting they cannot afford to close the reactors for lack of alternative energy supplies.

Thomas Land

SCIENTIFIC MISCONDUCT

NIH staffer finds warm reception on other side

Washington

ON the morning of 1 August, 1991 Jules Hallum, the director of the National Institute of Health (NIH) scientific misconduct office, and his deputy, Clyde Watkins, headed into Washington for what they knew would be a grim ordeal: a grilling by the staff investigators of Representative John Dingell (Democrat, Michigan).

But what promised to be a unsettling encounter turned out to be even worse. When the two officials arrived, they found themselves across the table from Suzanne Hadley, who had resigned as deputy director of Hallum's Office of Scientific Integrity (OSI) several months earlier. All morning, Hadley and the Dingell staff interrogated Hallum and Watkins about their misconduct policies. Doubly vexing for the two NIH officials was the fact that their former colleague was — and still is — an NIH employee; as far as NIH is concerned, Hadley holds a position in the agency's science education office, a job she took after being forced to leave OSI.

Since that first summer morning,

Hadley has worked for Dingell essentially at his will; with a simple telephone request from his oversight and investigations subcommittee, she switches from the Executive branch to the Legislative branch for a day. For Dingell, Hadley is a fortunate windfall: a scientific misconduct expert whose views are in conflict with those of NIH director Bernadine Healy. Since one of Dingell's chief targets is OSI itself (he believes the office is vulnerable to conflicts of interest under NIH control), its former deputy director is an invaluable witness for the prosecution — or a prosecutor herself.

For NIH, however, it is a more questionable call. When Dingell requested the services of Walter Stewart, an NIH scientist who has made a career as a freelance fraud-buster, NIH essentially transferred him to the subcommittee for most of a year, a not uncommon process in which federal employees are temporarily 'detailed' to Congress. Hadley, however, has not been detailed. She remains on the NIH payroll, even when she is investigating

the very agency that employs her, an apparent violation of the separation of powers between the branches of government required by the US Constitution.

NIH spokesman Don Ralbovsky says that Hadley is being made available to Dingell "primarily to answer questions". After several months of *ad hoc* arrangements, NIH and Dingell exchanged letters of agreement on Hadley's role last month, he says. In a statement released early this week, NIH stated that it had not been "aware that Dr Hadley was to participate in an interview of a [Public Health Service] employee being conducted by the subcommittee, nor had Dr Hadley been authorized to participate in any such interview." NIH policy is that it is not appropriate for NIH employees to "reveal or discuss without authorization confidential information obtained in the course of their current or previous duties," according to the statement. Hadley, NIH said, has "recently been reminded" of the policy.

Christopher Anderson

Universities win

Tokyo

THE Japanese government has decided to do what researchers worldwide wish their governments would do too — pour resources into university renovation. Japan's usually tight-fisted Ministry of Finance will invest approximately \$800 million in the renovation of major universities during the next five years. The sum, which far exceeds the \$600 million in renovation funds requested by the education ministry (Monbusho), shows the tremendous political importance that is now being attached to strengthening the basic research system.

Another sign of increasing interest in basic research is a decision to restructure Tokyo University's graduate school, with emphasis on mathematics, and on multidisciplinary research. At the end of August, Monbusho submitted a plan to the finance ministry to spend just over \$600 million on renovation over the next five years starting from fiscal 1992 (*Nature* **353**, 292; 26 September, 1991). Normally such budget requests are whittled down by the Ministry of Finance. But in late-night negotiations in the final days of last year, the finance ministry in an unprecedented decision agreed to increase the budget to \$800 million.

The decision follows a string of recent government advisory reports recommending more money for the universities and came shortly after the formation of a new political pressure group in the ruling Liberal Democratic Party which is pushing for more money for Japan's public sector research system (*Nature* **354**, 342; 5 December, 1991). The budget still has to be approved by the Diet, but approval is just a formality and no changes are expected.

Most of the money allocated for fiscal 1992 is expected to go to Tokyo University, Japan's leading national university. When the former minister of education Yutaka Inoue, visited Tokyo University last year he was appalled at the dilapidated state of the university and spurred moves for a drastic rebuilding programme. Next in line comes Kyoto University followed by other leading former Imperial universities, including Tohoku University.

In fiscal 1992, work will begin on a new ten-storey building for the faculty of engineering of Tokyo University, in 1993 the university's faculty of science will get a new building. The medical science faculty is also getting a new building this year under a separate budget.

Some of the money for the rebuilding programme will be raised by selling university land. There are, for example, plans to move some parts of Tokyo University to Kashiwa city in Chiba Prefecture east of Tokyo. Just which parts will be moved and which lands will be sold remains a

controversial issue to be resolved in the next year or so. But even small patches of land in Tokyo can be sold for astronomical prices, while prices in Chiba are comparatively low.

Also approved by the Ministry of Finance is a plan to reform the graduate school of Tokyo University. A new graduate school of mathematical sciences with 60 professors and assistant professors will be created by bringing together and upgrading the rank of staff members from the faculty of science and the college of arts and sciences. Tokyo is also in line for a new graduate school with giant multidisciplinary research groups comprised of the rest of the science faculty (*Nature* **351**, 679, 27 June, 1991). As a result, general research funds for each research group will increase by at least 25 per cent. The research group funds, which are automatically provided each year and are separate from competitive grants, are used for the general running costs of research. University researchers have been suffering from a severe shortage of such funds in recent years and the increase will

bring welcome relief to Tokyo University.

The faculty of engineering of Tokyo University will also undergo similar reform. But, because the number of staff members is much larger at about 200, reform will take several years. Also included in Monbusho's budget is a large 10 per cent increase in the budget for competitive grants, which are open to all university researchers, to \$520 million for fiscal 1992. The ministry has also won extra funds for scholarships and fellowships for graduate students. And a new programme of teaching assistantships for young researchers has been approved.

The huge influx of extra funds for the universities has wider political implications. Japan's bureaucrats and politicians have repeatedly told the United States that they cannot provide money for the US Superconducting Super Collider (SSC) until the situation in the universities is improved. The extra funds that have been approved for the universities will make it easier for Prime Minister Kiichi Miyazawa to offer financial support for the SSC when President George Bush visits Tokyo this week.

David Swinbanks

MITI Science and Technology Budget for 1992

	thousand million yen	% change from 1991
Total R & D budget	259.7	+ 1.5
Japan Key Technology Centre	26.0*	- 9.7
Basic technologies for future industries	8.2	+ 5.6
Large-scale industrial projects	14.7	+ 4.8
Sunshine new energy sources project	26.5	+ 8.6
Moonlight energy-saving project	11.8	+ 4.7
Fifth-generation computer	3.6	-49.7
Sixth-generation computer	0.9	+842.7
Global environment	8.2	+ 22.1
Intelligent Manufacturing System (IMS) project	0.8	+197.8
Human Frontier Science Program (HFSP)	3.8**	+ 3.5
Elucidation of biological functions (domestic HFSP)	0.3	-12.8
NEDO international grants	0.7	+53.4
Reorganization of Tsukuba science institutes	14.9	+11.3

* budget shared with Ministry of Posts and Telecommunications

**budget shared with Science and Technology Agency

Science and Technology Agency Budget for 1992

	thousand million yen	% change from 1991
Total R & D budget	551.8	+ 5.7
Special Promotion Funds	11.0	+ 4.8
Space	144.7	+ 9.9
Nuclear Energy	315.2	+ 2.9
SOR (Spring-8)	7.0	+43.0
Ocean Research	11.4	+ 6.8
ERATO	6.3	+12.1
Sakigake	1.0	+117.0
Human Genome	1.1	+20.9

Tokyo

Apart from the education ministry, there are no particular surprises in the budgets for Japan's science-related ministries and agencies which were approved by the Cabinet on 28 December. Of note, the Science and Technology Agency gets a 10 per cent boost in its space budget, much of it for Japan's contribution to the US space station (28,200 million yen). A similar amount is allocated for final development of Japan's new H-II rocket, which should be launched in 1993.

AFRC lays down the law

London

BRITISH university departments and laboratories run by the Agricultural and Food Research Council (AFRC) are being sent copies of the AFRC's new guidelines on the use of animals in research. This is the first time that one of the research councils has released its own document to ensure that animal researchers are fully aware of their responsibilities under the law. AFRC funding will in future be conditional on researchers having read and complied with the new guidelines.

The guidelines themselves are uncontentious, merely summarizing Home Office requirements to comply with UK animal experimentation legislation, and the separate agriculture ministry rules for the use of farm animals in research. But animal welfare groups say their release is significant, marking a new realization among British research funding agencies that ensuring animal welfare should not simply be left to government inspectors responsible for policing the law.

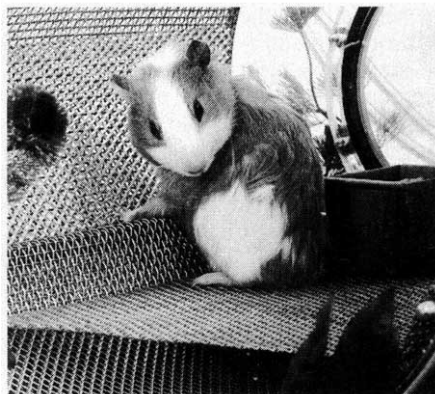
Michael Balls, director of the Fund for the Replacement of Animals in Medical Experiments, believes the AFRC's move is, in part, a response to the infamous 1990 'Feldberg case', where the Scottish animal welfare group Advocates for Animals exposed serious breaches of the 1986 Animals (Experimental Procedures) Act by Wilhelm Feldberg, a researcher at the Medical Research Council's (MRC) National Institute for Medical Research in London.

Mari Williams, secretary to the AFRC working party that drew up the new guidelines, says that the AFRC group was intending to set up an animal research working group before the Feldberg revelations surfaced, but admits that the incident "focused minds". The MRC itself has not issued a similar document, but did set up a new secretariat to monitor animal research in the wake of the Feldberg allegations.

Meanwhile, a second Advocates for Animals investigation has been generating considerable heat. In November, the group released the first instalment of a three-part report detailing primate research papers which, it was alleged, should not have been approved under the 1986 Act (see *Nature* 354, 177; 21 November 1991). The MRC, which funded more than half of the projects involved, has since published its own internal inquiry into the allegations, concluding that in each case the work was directly relevant to human medical disorders and that the results could not have been achieved without the use of animals.

But the Advocates for Animals report now seems to have been picked up by the extremist wing of the British animal rights

movement. Shortly before Christmas, *The Scotsman* newspaper received a letter from the militant Animal Liberation Front (ALF), claiming responsibility for a recent spate of firebomb attacks against Scottish research laboratories and meat-handling depots, and threatening to destroy the primate unit at the MRC's Reproductive



LONDON ZOO

Rebel fellows claim success

London

DISGRUNTLED fellows of the Zoological Society of London (ZSL) who have been seeking to replace the management team responsible for running the society's showcase, London Zoo, claimed a major victory on Monday this week after a meeting called to debate the ZSL's attempts to address the financial crisis that has threatened the zoo with closure. The meeting was called by the *Ad Hoc* Group of ZSL Fellows, to debate a motion of no confidence in the ZSL's governing council, and to secure changes in the zoo's management (see *Nature* 354, 261; 28 November 1991).

Although the council was saved the indignity of being forced to resign, Graham Mitchell, from the medical school at Guy's Hospital, London, and a member of the *Ad Hoc* Group, believes the group's action has forced the council to rethink its approach — and that there is now cause for optimism about the zoo's future. He argues that the council has now rejected the zoo management's "grandiose" plans to redevelop the zoo's Regent's Park site into a centre featuring large 'themed' exhibits featuring the ZSL's work in conservation. In their place, says Mitchell, the plan now given the council's backing includes many of the features of the *Ad Hoc* Group's own low-budget proposal, which featured captive breeding programmes and relied on the re-use of existing buildings, rather than expensive redevelopment.

The sequence of votes taken at the

Biology Unit in Edinburgh, if it is not closed within 16 weeks. Three of the 13 papers detailed in the first part of the Advocates for Animals report were by researchers at the Edinburgh unit, looking at reproductive behaviour or the secretion of reproductive hormones, and involved either surgery or aggression-induced stress.

Advocates for Animals and the MRC have since released a joint statement condemning the ALF threat, and Advocates for Animals director Les Ward says that he is now looking again at how best to present the remaining two parts of his group's primate research investigation, due to be published later this month. The first part of the report named the researchers and laboratories involved in the work. Ward says he is considering omitting this information from the rest of the report, although he thinks that the ALF threat is probably a hoax, as he doubts if the group would attempt to destroy a laboratory that is still housing animals.

Peter Aldhous

meeting yields a confusing picture: fellows declared their lack of confidence in the council's past performance by 208 votes to 56; but the council gained the backing of the same gathering to continue in its attempts to save the zoo by 141 votes to 96 (with an increased number of abstentions). But Mitchell points out that the second vote came after the presentation of the council's new redevelopment plan, costed at around £9 million — £7 million less than the least expensive figure quoted before Christmas, and a fraction of a mammoth £60 million proposal touted by a specially created private company, Regent's Park Zoo Limited.

After the council's treasurer, Peter Holwell, assured fellows that the zoo's management would also be overhauled (the *Ad Hoc* group has been highly critical of the zoo management, which it regards as ineffective and 'top heavy'), the rebel fellows pulled back from their threat to remove the council, with the request for the council to resign being defeated by an overwhelming margin.

Despite the renewed optimism among the council's critics, there is now very little time in which to assemble financial backing to secure the zoo's future, with closure threatened sometime later this year if no firm offers materialize. Mitchell says that the council must act on its promise to reform the zoo's management — something that may place the jobs of some senior staff in jeopardy — if further challenges from the fellowship are to be prevented.

Peter Aldhous

Writer's cramp

Washington

WHILE most researchers struggle to publish a few papers a year, Yury Struchkov gets his name in print almost twice a week. Over the past ten years, the Russian chemist has been listed as an author on almost a thousand scientific articles.

Impossible? Not, apparently, if one runs a big crystallography laboratory in Moscow. A few days' work by researchers in Struchkov's group at the Institute for Organoelemental Chemistry is usually enough to generate a paper, and Struchkov is an author on every one. A new survey of extraordinarily prolific researchers, which ranks Struchkov number one worldwide, raises once again the question of who should be an author and who should not.

Twenty researchers worldwide have published an article at least once every 11.3 days over the past decade, according to a report released last week by the Philadelphia-based Institute for Scientific Information (ISI). The top five researchers have published more than once a week. Some of these authors, like Struchkov, are crystallographers who, by nature of their sophisticated equipment and experiences, collaborate with dozens of researchers who need their services. Co-authorship is usually given in return.

Most of the others are scientists who run medium-sized to large laboratories or research groups that have tapped into a particularly fruitful line of research in their field. In the top twenty, the single discipline most heavily represented is basic molecular chemistry, followed by transplant surgery. Biomedical research, in general, accounts for more than half of the total.

An informal survey of many of the researchers on the list reveals that there are as many authorship policies as there are authors. David Greenblatt, a Tufts University pharmacologist, insists on full participation in his research. "I do the work", he says, "with my own two hands." Others, like John Najarian, a transplant surgeon at the University of Minnesota, take a more hands-off approach. He contributes ideas, advice and reviews the paper written by researchers in his department.

Both scientists put their names on almost all the papers that come out of their research groups. But as science finds itself under scrutiny as never before, many researchers are reconsidering the old problem of authorship. Is the loan of a key reagent worth a co-authorship? How about a good idea? Regular guidance?

As a result of cases in which prominent researchers were damaged by the revelation that papers on which they were listed as authors contained fabricated data (even though they had not done the research themselves), the debate over

authorship is no longer academic. A laboratory director who initiates or supervises a project is now considered responsible for the accuracy of the data itself if he or she shares authorship.

Last November, after receiving a paper with more than 200 co-authors (including departmental secretaries), the *New England Journal of Medicine* announced new authorship policies. It would henceforth require that anyone designated an author make "substantial contributions" to three elements of the research: conception, design, or analysis and interpretation of the experiment; drafting or critically revising the article; and reviewing and approving the final draft.

Harvard is another institution that is worried about publication practices. In an effort to reduce the emphasis on

to vary by discipline. Because chemistry experiments are relatively straight-forward and self-checking, the heads of prolific chemistry laboratories say they feel confident in simply providing ideas and oversight, along with regular review, to work they will eventually co-sign. "I generate the ideas and I direct the research," says Alan Katritzky, a University of Florida chemist who co-authors all the papers that come from his laboratory of 35 to 40 researchers. Although he examines primary data in weekly meetings with his team, it is not generally to check for fraud. "It's very difficult to fudge data in organic chemistry," he says. Frank Cotton, a Texas A&M University chemist, says he selects — and co-authors — most of the research projects in his laboratory because the grants are in his name, and are based on his proposals. Although he does little of the bench research, he says he guides his 15 to 25 researchers, examines their data, and

World's twenty most prolific researchers

	Name/Field/Nation	No. papers* 1981-90	Ave. days per paper	Ave. citations per paper
1	Yury Struchkov/Chemistry/USSR	948	3.9	3.0
2	Stephen Bloom/Gastroenterology/UK	773	4.7	21.4
3	Mikhail Voronkov/Chemistry/USSR	711	5.1	2.0
4	Aleksandr Prokhorov/Physics/USSR	589	6.2	3.1
5	Ferdinand Bohlmann/Chemistry/Germany	572	6.4	6.2
6	Thomas Starzl/Surgery/USA	503	7.3	16.8
7	Frank Cotton/Chemistry/USA	451	8.1	11.4
8	Julia Polak/Histochemistry/UK	436	8.4	26.6
9	Robert Gallo/Cell Biology/USA	428	8.5	86.0
10	Genrikh Tolstikov/Chemistry/USSR	427	8.5	1.2
11	John Huffman/Crystallography/USA	403	9.1	13.2
12	Alan Katritzky/Chemistry/USA	403	9.1	4.5
13	David Greenblatt/Pharmacology/USA	383	9.5	17.1
14	John Najarian/Surgery/USA	345	10.6	14.6
15	Willy Jean Malaisse/Endocrinology/Belgium	344	10.6	10.9
16	Charles Marsden/Neurology/UK	339	10.8	15.0
17	Anthony Fauci/Immunology/USA	338	10.8	52.5
18	E. Donnall Thomas/Oncology/USA	328	11.1	37.5
19	Noboru Yanaihara/Biochemistry/Japan	322	11.3	14.0
20	Timothy Peters/Biochemistry/UK	322	11.3	9.5

Source: ISI's Science Indicators Database 1981-90.

* papers defined as articles, reviews, notes and proceeding papers; abstracts, letters, corrections, etc. were not counted.

publication volume in science, the University now asks would-be professors to select only their top ten papers for tenure consideration.

Debate over authorship is nothing new. The senior scientist who puts his name on every paper in his laboratory has become a virtual stereotype. At the Russian Institute of Volcanic Geology and Geochemistry earlier this year, ten geologists actually went on a hunger strike to protest a "autocratic" director who forced institute researchers to list him as a co-author on their papers (see *Nature* 354, 3; 1991). On the other side is the ambitious young researcher who appends the names of distinguished advisers to his papers to improve their chance of publication. Since becoming well-known, "I've been taking my name off more papers than ever," says Anthony Fauci, of the US National Institutes of Health.

Differences in authorship policies tend

has a hand in writing almost all the papers.

"I only put my name on if I had the idea, started the study, or played an active part in it," says Julia Polak, a University of London pathologist. "My own criteria is that if I don't understand it and haven't been part of the writing of the paper, I don't want my name on it." After watching the trials of Nobel Laureate David Baltimore, who unwisely defended a 1986 *Cell* paper on which he was a co-author although he had done none of the original research, Polak has instituted new data policies in her laboratory. Researchers now bring their primary data to weekly meetings, and archive even secondary data in case it is ever challenged. Although there may never be firm, interdisciplinary rules about authorship, prolific researchers do seem to be aware now of the perils of appending their names to work they have not closely supervised.

Christopher Anderson

Confusion over therapy

London

CONFUSION generated by the halting of a trial of the antiviral compound acyclovir in HIV-infected patients has embarrassed the drug's manufacturers, the British company Wellcome — not to mention some sections of the British press.

On 29 December, the lead story in *The Sunday Times* proclaimed a major breakthrough in AIDS research, saying that a trial of the drug in 300 patients in Britain, Germany and Australia had revealed that the death rate in those receiving acyclovir and Wellcome's zidovudine (AZT), the main anti-retroviral agent used to combat HIV, was half that in the group receiving AZT alone.

The article reported Paul Griffiths, a virologist at the Royal Free Hospital in London involved in the trial, as saying that AIDS could become as treatable as diabetes by the end of the decade, and stated that the trial had been stopped on ethical grounds to allow patients in the control group access to acyclovir.

But a slightly different story has since emerged from Wellcome, the trial's sponsors. Paul Fidian, Wellcome's head of clinical virology, says that the finding of increased survival in patients given both drugs is not new, and that the trial reported in *The Sunday Times* was halted because preliminary data had revealed that orally-administered acyclovir does not seem to be effective in AIDS patients against cytomegalovirus (CMV), one of the herpes viruses.

The earlier trial showing an increase in survival after 'combined therapy' with AZT and acyclovir was also funded by Wellcome. Its results were presented — with little fanfare — at the 1990 international AIDS conference in San Francisco, although a paper based around these results has not yet been published. Fidian says that the working hypothesis at the time was that acyclovir, a known anti-

herpes agent, had improved survival by acting against CMV, which was considered the herpes virus most likely to be fatal in AIDS patients. With this in mind, a second trial was designed to look more closely at the effects of acyclovir on CMV in HIV-infected patients whose CD4 cell counts had been reduced to less than 150.

It is this trial which was halted by Wellcome shortly before Christmas, when Fidian and his colleagues found no indication from preliminary data that acyclovir was effective against CMV. Wellcome is now acutely embarrassed by the apparently contradictory stories.

Spokeswoman Rosemary Hennings says that the company is concerned about the effect the coverage will have on HIV-infected people, who "must be spinning at a rate of knots" in confusion. But one researcher involved in the trial criticizes Wellcome for announcing that the trial had been stopped just before the Christmas break, when few staff were available to field journalists' questions. Griffiths, whose conversation with *The Sunday Times* sparked the controversy, was earlier this week unwell and not available for comment.

The trial's inconclusive result also leaves a perplexing question: what exactly is causing the increased survival reported in patients receiving both AZT and acyclovir? Fidian presumes that the effect is related to the action of acyclovir against one or more of the herpes viruses, although the precise mechanism is still a mystery.

Wellcome has also not ruled out using acyclovir-like compounds to combat CMV in AIDS patients. One possibility being investigated is to give an oral dose of a precursor compound that will be absorbed rapidly, and converted to acyclovir once in the body — giving a much higher blood concentration of acyclovir than is possible by giving the drug itself orally.

Peter Aldhous

TECHNOLOGY

Staff buyout for BTG?

London

THE management and staff of the British Technology Group (BTG), one of the world's principal technology transfer organization, are leading one of several bids to buy the company, when it is sold into the private sector later this year. Since legislation to privatize the state-owned company completed its passage through parliament last October, interested consortia have been submitting their bids to the accountants Price Waterhouse, who are handling the sale for the government.

Tony Christmas, BTG's head of marketing, says that the consortium put

together by management and staff comprises a number of investment trusts, foundations and pension funds (the government has said that no single shareholder should own more than 15 per cent of BTG's shares), but declines to name the parties involved. Price Waterhouse are similarly coy about the number of bids received, although Christmas says "we think we know of two or three others". The accountants have yet to place a value on BTG, but the company is thought to be worth around £40 million. BTG hopes to begin detailed negotiations soon.

Peter Aldhous

Explosion kills one

ONE researcher was killed and three injured last week in a chemical explosion at one of last US laboratories studying 'cold fusion'. Officials at the Stanford Research Institute (SRI), site of the blast, believe that a buildup of either deuterium or hydrogen in a six-inch tall reactor chamber caused the apparatus to explode, killing Andrew Riley, an Oxford University-trained researcher on contract with the Electric Power Research Institute (EPRI). Michael McKubre, who directed the SRI project, Stuart Smedely, another SRI researcher, and Steven Crouch-Baker of EPRI were injured by flying glass and metal. SRI checked and found no radiation after the blast, but has buried the remaining three reactor chambers until a investigation determines the exact cause of the explosion. Menlo Park, California, fire department authorities said that a pressure valve in one canister had stuck and that the researcher were attempting to manually release pressure when the container burst.

The SRI researchers were looking for excess heat production in an experimental reactor similar to that first used by Stanley Pons and Martin Fleischmann, the two University of Utah electrochemists who announced in 1989 that they had discovered a new energy-generating effect. The small reactors contain palladium electrodes in a deuterium bath, or a control set-up replacing the deuterium with regular water. After passing a current through the reactor, the Utah researchers claimed to detect an inexplicable excess of heat. For the past two years EPRI has been funding additional studies on the phenomena, trying to find an explanation for measurements of anomalous low-level excesses in both heat and nuclear product production. EPRI has spent \$2 million on the SRI project to date.

Christopher Anderson

US ACADEMY OF SCIENCES

IOM gets top name

Washington

KENNETH I. Shine, dean of the University of California, Los Angeles, School of Medicine, will become the next president of the Institute of Medicine (IOM). A heart specialist, Shine has been a long-time campaigner for hospital care reform. He will replace Samuel O. Thier, who left IOM last September to become president of Brandeis University. Shine will be the sixth president of the institute, which was created in 1970 as the medical policy arm of the National Academy of Sciences. Shine, a graduate of Harvard Medical School, will assume the Washington-based IOM presidency in July.

Christopher Anderson

Nature of human sexuality

SIR — Nessa Carey (*Nature* 354, 9; 1991) is right to say that homosexuals have been given a bad time by sections of the Church and the medical profession. AIDS is a terrible disease, but it would be foolish indeed not to take advantage of the perhaps unique opportunity it has presented to probe into the interaction between a virus and the immune system. This has resulted in an explosion of interest and research in the scientific community. Among the topics brought into focus is the nature of human sexuality itself. Discussions appear in the press and media that would have been unthinkable only a few years ago.

Human sexuality has always fascinated mankind, as is evident by the many 'Venuses' and phallic objects that have been found from the Stone Age onwards. Plato in his *Symposium* describes three sexes, males, females and androgynes, each split in half by an angry Zeus, and each forever seeking its partner — a neat way to explain heterosexual and homosexual attraction. There has been no shortage of theories since, but it was Kinsey¹ who in his famous 0-to-6 scale described humanity not as sheep or goats but with some 50 per cent of humanity at the heterosexual end (0,1) of the scale, 5 per cent at the exclusively homosexual end (5, 6) and the rest in between.

More elaborate and useful descriptions have since been propounded² which further differentiate between biological sex, sexual identity, sexual behaviour and preferred sexual outlets and so on, and how these change with time. Some have argued that homosexuality as a fixed entity across culture and time does not exist, but that homosexuality itself is merely a social construct³, while others believe in a homosexual 'essence' and cite common factors in the lifestyles and behaviour of homosexuals across culture as evidence⁴.

Dawkins⁵, has suggested that even if there are homosexual genes in modern man, in earlier times, say in our Pleistocene ancestors, these may have interacted with other genes and environments to produce completely different phenotypic characteristics and behaviour, which may have had nothing to do with homosexuality. This is perhaps why efforts to understand the role of genes and intrauterine hormones in the determination of sexual variation in man have proved so unrewarding.

Understanding the role of the hypothalamus, and indeed the brain as a whole, in the construction of gender identity and sexuality is of fundamental importance, and Simon LeVay's work may just offer a key to the process. A luddite approach will help nobody. A

biological basis for homosexuality would suggest that it is a part of normal human sexuality, and should no longer be regarded as a mental illness or as an aberration.

R. E. GOODMAN

Psychosexual Clinic, Lancastrian Unit, Hope Hospital, Salford M6 8HD, UK

1. Kinsey, A. C., Pomeroy, W. B. & Martin, C. E. *Sexual Behavior in the Human Male* (Saunders, Philadelphia, 1948).
2. Coleman, E. J. *Homosex*. 14, 9–24 (1987).
3. Weeks, J. *Sexuality and Its Discontents* (Routledge, London, 1985).
4. Whitham, F. L. *Archs sex. Behav.* 12, 207–226, (1983).
5. Dawkins, R., *The Extended Phenotype* (Oxford University Press, Oxford, (1982)).

Consistent policy?

SIR — Commercialization of biological sequence data is certainly worrisome and we were pleased to see the attention given to the subject in *Nature* (354, 171; 1991). For many years the nucleotide sequence databases have been encouraging authors to submit their data to the appropriate database and working with journals to ensure that authors transfer data. Most molecular biology journals now require that sequence data reported in a publication be deposited in the relevant database. *Nature* is the exception. It was, therefore, something of a surprise to read that *Nature* is now strongly in favour of sequence deposition in databanks.

On the one hand, the leading article states: "Both MRC and NIH are thus looking for speculative (but probably insubstantial) benefits from the first flush of data with a new technique. The policies are discreditable because they make a monkey of honest research, and because so little thought has been given to the damage they will do. What, for example, will happen to exhortations of those in the academic sector with novel sequence data that they should add them to the public databanks? The case for doing so is strong, for these are eminently databanks in which the whole is greater than the sum of the parts."

On the other hand, *Nature* as a scientific journal does not think the matter serious enough to warrant more than a mere request to authors: "*Nature* requests authors to deposit sequence and crystallographic data in the databases that exist for that purpose, and to mention the availability of these data." (Guide to authors, same issue).

One of the main reasons that progress has been so rapid in molecular biology is the free and rapid exchange of data and materials (clones and so on). It is in nobody's interest — not even industry's — to see this cease. *Nature* and other journals can play an important role by

insisting that all data be deposited in machine-readable form in the public databases. Similarly, granting agencies can make it clear that grants are awarded on the understanding that data will be deposited.

PETER R. SIBBALD

LENNART PHILIPSON

GRAHAM CAMERON

European Molecular Biology Laboratory, Postfach 10.2209, W-6900 Heidelberg, Germany

■ There is no inconsistency. *Nature* shares the general view that the databases are invaluable. One of the many reasons why it does not require that data should be deposited in them is that their managers have given too little thought to their commercial use. It is a pity that the authors of this letter have not dealt with that issue, now even more topical. — Editor, *Nature*.

Burt files reopened

SIR — Although it is much to the credit of the British Psychological Society (BPS) to be reconsidering its public discrediting in 1980 of the late Sir Cyril Burt (see *Nature* 354, 517; 1991), since it has itself been very much *parti pris* on the question, the BPS cannot now be expected to conduct a truly impartial inquiry into whether it was wrong in what it did ten years and more ago.

As I suggested in an article in the *Oxford Magazine* (No. 70) last May (my efforts over the previous year to raise the matter in more widely circulated publications having been frustrated!), this would better be done by the British Academy, of which Burt had been a highly respected fellow. The charges of fabrication of data and scientific fraud, first made shortly after his death in 1971, were, however, so serious that they would almost certainly have resulted in his expulsion if they had been substantiated during his lifetime. The British Academy should therefore still have a proper interest in determining whether the charges were indeed well founded.

It is not a question of Burt's rather unfashionable hereditarian views on intelligence and educability being correct, let alone politically correct, but whether he arrived at them dishonestly, by fraud and fabrication. The British Academy has plenty of fellows well able to judge of this, which requires no expertise in educational psychology (better not, in fact, for an unprejudiced point of view), and a small group of them could be asked to examine the evidence and to report as to whether a serious injustice has been done to the posthumous reputation of someone who was once one of their more distinguished fellows.

C. B. GOODHART

Gonville and Caius College, Cambridge, CB2 1TA, UK

Patenting of genes

SIR — As a patent attorney and Human Genome Project enthusiast, I believe that the lead-in to your recent report¹ on Craig Venter's patenting of uncharacterized cDNAs deserves some clarification. Patenting genes relating to genetic disorders such as cystic fibrosis (CF), Huntington's disease (HD) and Duchenne muscular dystrophy (DMD) is not, as stated in your article, "business as usual". This is illustrated by a preliminary patent search, summarized in the table, directed towards these disorders.

This table indicates, first, that there are numerous, potentially conflicting, patent applications being filed in this area. This is frequently true in the case of biotechnology patents. Second, the patent applications listed often claim methods of disease detection that are not limited to specific probes or genes. This is a key strategy for obtaining commercially significant patent protection. Third, and most importantly, these patent applications overlie one of the primary goals of the Human Genome Project: the advancement of testing and possible treatment for genetic diseases².

One may argue about whether or not the patent system will further this goal. It is undeniable, however, that the issue of "who owns the human genome" has been a subject of controversy since the early days of the genome project, when copyrights for genetic maps and sequences were being claimed³. Venter's patent application has served merely to reignite the controversy.

In addition to the listed patent applications, patent applications relating to genes associated with various cancers, atherosclerosis, familial hypercholesterolaemia and Alzheimer's disease have also been published. Even though Ven-

ter is himself reported to have said that researchers would naturally want to take time to characterize sequences before publishing them⁴, the National Institutes of Health attorney in Venter's case, aware of this patent activity, must respond to the pressure to file early and often.

As more DNA is sequenced, more markers are discovered and more mapping is completed, it will become more urgent that the scientific (and business) community come to a clear understanding of what will — and will not — be in the public domain as a result of the Human Genome Project.

DAVID J. ASTON

1689 Sunnyslope Drive,
Belmont, California 94002, USA

SIR — A group of research workers wishes to patent a ragbag assortment of anonymous human gene sequences derived from a cDNA library of brain tissue (*Nature* **353**, 485 & 785; 1991). I find this ludicrous, and suggestive of a lack of ethical and professional judgement.

Was specific informed consent given by the owner (or former user) of the brain tissue used to construct the cDNA library? Did that person agree that the researchers could profit from his or her death, or was the donation given for the advancement of science and medicine rather than a few bank balances?

In fact, can the researchers really claim to have invented these sequences? Is not the real inventor the individual from whom they are derived, or his/her parents, grandparents and on back down the lineages to Olduvai and Lucy? Maybe we should credit natural selection, genetic mutation and random drift

with the invention. Or, following the arguments of Richard Dawkins in *The Extended Phenotype* (Oxford University Press, 1983), we could regard the true inventor of the sequences to be the sequences themselves, with the individual being the organismal carrier they invented to propagate themselves.

On the matter of prior publication, the patent claim relies on the hope that these sequences have been published previously — in hundreds or thousands of copies in thousands or millions of brain cells in millions or billions of individuals for a fair chunk of evolutionary history. If they were not so expressed they would be of no interest, commercially or medically. I believe that these basic sequence data are a heritage we all share, and should share. Just because the book of life has not been written in human language does not give us the right to plagiarize (and patent) its contents.

MARK BLAXTER

Wellcome Research Centre
for Parasitic Infections,
Imperial College,
London SW7 2BB, UK

1. Anderson, C. *Nature* **353**, 485–486 (1991).
2. Watson, J. *Science* **248**, 44–49 (1990).
3. Roberts, L. *Science* **237**, 368–361 (1987).
4. Roberts, L. *Science* **248**, 953 (1990).

Page charges

SIR — Publication charges are imposed on authors by some journals for (1) processing the manuscript for reviewing, (2) publication cost of each page and (3) the cost of reprints. Why do only some journals do this? Some authors do not have grants and even those who have grants may be reluctant to submit their papers to such journals because of strict regulations on foreign exchange; procuring even small amounts of foreign currency may be time-consuming, and no scientist would want to delay publication of his data.

Could not such journals waive page charges, at least for authors based in countries where there are strict foreign exchange regulations? Alternatively, could they not accept payment of the charges in any currency? They should at least remove the processing charges, which would provide an equal opportunity for all contributions, irrespective of the country of origin, to have their work critically reviewed by their peers.

Y. D. SHARMA

Department of Biotechnology,
All India Institute of Medical Sciences,
New Delhi-110029, India

Letters submitted for Correspondence should be typed, double-spaced, on one side of the paper only.

PATENTS RELATED TO GENETIC DISORDER

Serial number	Institution	Subject matter
EP 226 288	Collaborative Research Corporation	Testing for CF allele
EP 288 299	St Mary's Hospital Medical School, UK School, UK	Diagnosis of CF with specified genomic DNA
EP 446 017	Genzyme Corporation	DNA sequence comprising CF transmembrane conductance regulator
WO 91/10734	HSC Research Development Corporation, Toronto	CF DNA with specific mutation
WO 90/15880	USA	Cell line transfected with CF gene
WO 88/03572	John Radcliffe Hospital, UK	DNA sequence deleted in DMD
WO 88/00979	Children's Medical Centre, Boston, MA	Detection of DMD using specified polymorphism
U.S. 4,666,828	General Hospital Corporation, Boston MA	Test for HD, based on polymorphism
WO 91/05044	Imperial Cancer Research	Yeast artificial chromosome containing mutant HD gene

Sharing out NASA's spoils

Roger H. Bezdek and Robert M. Wendling

The economic benefits of NASA's programmes are greater than generally recognized. The main beneficiaries may not even realize the source of their good fortune.

THE US space programme is mired in controversy concerning the future of the manned space station, the role of the space shuttle versus expendable launch vehicles, the priority of basic science programmes within the National Aeronautics and Space Administration (NASA), and problems of quality control and cost overruns. The wisdom of spending huge sums on projects such as the space station is questioned by many scientists, who are concerned that there may not be enough money for basic research, and by elected representatives, who would like to use the money for earthly priorities. Inevitably, advocates of the space programme revert to economic arguments. NASA, for example, emphasizes that the space station programme alone will provide orders for 2,000 businesses and create thousands of jobs in 40 states¹.

Such rationalizations are disingenuous and incomplete. Any large capital-intensive project involving billions of dollars per year will create profits for many businesses and jobs for many workers. More interesting, though, is the fact that advocates of the space programme ignore many of its economic benefits: the indirect, pervasive effects of NASA expenditure throughout the economy.

These benefits are distinct from the intangible rewards that analysts identify (spin-offs, research and development support, public goods, and creation of new space industries) for the following reasons. (1) Spin-offs are products developed for the space programme that have applications in other areas, including photovoltaics, aerodynamic design, telecommunications systems, microelectronics, chemical processes and various consumer-orientated applications². (2) The space programme contributes to research and development in general and is important for the technological competitiveness of US industry; investments in NASA research and development have a high rate of return³. (3) Public goods are those that only the government can provide, as their benefits cannot be captured by private investors, for example knowledge about the Universe, information on the characteristics of the Earth and Solar System, and related basic scientific knowledge⁴. (4) New space-based industries include private launch services, materials processing in space and related applications of micro-

gravity, remote imaging, infrastructure development and so on: without a strong space programme, development of these may be delayed and opportunities lost⁵. There are more immediate, quantifiable economic benefits of the space programme: sales, profits, jobs and tax revenues.

State benefits

In 1987, NASA procurement spending was \$8,600 million; 70 per cent of prime contract awards were concentrated in California (the centre of the US aerospace industry), Texas (location of the Johnson Space Center), Florida (home of the Kennedy Space Center), Maryland and Alabama⁶. This corresponds to the conventional wisdom that the economic benefits of the space programme are concentrated within a few industries in a few states⁷.

But we find that the economic benefits of the space programme are much more widespread than has previously been realized. Specifically, in 1987, the NASA procurement budget generated \$17,800 million in total industry sales, had a 'multiplier effect' on the economy of 2.1, created 209,000 private-sector jobs and \$2,900 million in business profits, and generated \$5,600 million in federal, state and local government tax revenues. Furthermore, these benefits are widely distributed throughout the country; among the biggest state 'winners' are many that are not perceived as being closely tied to such expenditures.

Why is the conventional wisdom so wrong? Before examining the answer to this question in detail, we illustrate the concept of indirect economic benefits by a simple example. California is a significant benefactor of NASA expenditures because of the concentration of the aerospace industry around Los Angeles. But contracts awarded in Los Angeles for space hardware require systems, components, materials and so on from industries in other states. These companies in turn require components, products and subsystems from industries located elsewhere. Many of the businesses involved are not even aware that they are part of the US space 'industry'. For example, businesses producing the wiring, paint or valves necessary to satisfy the third or fourth round of indirect output requirements usually have no idea that their sales, profits and jobs are being generated by the space

programme.

NASA can trace its main contractors and their subcontractors, but it is impossible to use accounting methods to trace the further impact of spending. In effect, NASA has never been able rigorously to identify even half of the economic effects of its programmes. If NASA is unable to determine the complete impact of its programmes, how could we? We used the MISI database and modelling system, which is based on economic input-output analysis, a quantitative method of measuring and analysing the interdependence of all industries in an economy.

Input-output analysis was developed by Leontief⁸, and we have used it to investigate the impact of the US space programme on the economy. First, we translated expenditure for NASA programmes into per unit output requirements for every industry in the economy. This is determined by four main factors: the state of technology, the distribution of expenditures, the specific programme configuration and the direct industry requirements structure. Although the MISI system contains 500 industries, we used an 80-order industry scheme.

We next estimated the direct output requirements of every affected industry. These direct requirements show, proportionately, how much an industry must purchase from every other industry to produce one unit of output. Direct requirements give rise to subsequent rounds of indirect requirements. The sum of the direct plus the indirect requirements represents the total output requirements from an industry necessary to produce one unit of output. Input-output techniques allow direct as well as indirect production requirements to be computed, and these total requirements are represented by the 'inverse' equations in the model. These equations show not only the direct requirements, but also the second, third, fourth, n th round indirect industry and service-sector requirements resulting from expenditure on the space programme. The ratio of the total requirements to the direct requirements is called the input-output multiplier.

Thus, in the third step, we convert the direct industry output requirements into total output requirements for every industry by means of the input-output inverse equations, and we use the total

TOTAL SALES, TOTAL EMPLOYMENT AND INDIRECT MULTIPLIERS RESULTING FROM 1987 NASA PROCUREMENT

Industry	Sales (\$ million)	Jobs (number)	Multiplier*
Electronic components	503.3	7,906	5.9
Electric lighting and wiring equipment	43.3	552	4.8
Electric, gas and sanitary services	612.3	2,108	4.5
Chemicals and selected chemical products	279.8	1,521	3.3
Wholesale and retail trade	668.6	20,705	3.2
Paints and allied products	26.2	177	2.9
Glass and glass products	24.6	312	2.8
Communications, ex. radio and TV	206.6	2,181	2.7
Business services	923.6	19,013	2.4
Transport and warehousing	785.4	10,832	2.3
Electrical transmission equipment	151.8	2,324	2.2
Average, all industries	—	—	2.1
Petroleum refining and related industries	781.0	815	2.0
Motor vehicles and equipment	250.1	1,233	1.8
Metal containers	15.0	84	1.7
Engines and turbines	120.4	1,065	1.4
Aircraft and parts	2,827.1	27,217	1.3
Miscellaneous transport equipment	375.3	5,002	1.1

This table refers only to selected industries. Additional data available on request from the authors.

*Ratio of total to direct output requirements.

output requirements from each industry to compute sales volumes, profits and value added for each industry. Then, using data on person hours, labour needs and productivity, we estimate employment requirements in each industry.

Finally, the total number of jobs in each industry is translated into occupational employment requirements using databases showing the occupational distribution of employment in each industry. We used data on the occupational composition of the labour force and estimated job requirements for 475 occupations encompassing the entire US labour force. This allows an estimate of the impact of the programme on jobs for specific occupations and on skills, education and training requirements.

This modelling approach allowed us to estimate the effects on employment, personal income, corporate sales and profits, and government tax revenues in the United States and in each state. Estimates were then developed for detailed industries and occupations. We simulated the effects of only the procurement portion of the NASA budget and, as we wished to analyse the impact on the private sector, we did not include NASA federal civil service employees.

Our results show that five states — California, Texas, Florida, Maryland and Alabama — receive 70 per cent of NASA procurement spending. But although this seems to illustrate that the economic benefits of the space programme flow primarily to a few regions, major (indirect) benefactors of the space programme include New York, Illinois, Michigan, Pennsylvania, Indiana, Missouri, New Jersey and Wisconsin. These states represent the manufacturing heartland of the country. Georgia, Mas-

sachusetts, Washington, North Carolina and Tennessee also benefit substantially.

Each of these states has a multiplier well above the national average of 2.1, and the multipliers for several states exceed 10 to 1. For example, for every dollar Indiana receives directly in space programme funds, is also receives \$12 indirectly in business arising from the programme. Similarly, Illinois receives \$8 indirectly per direct dollar spent, Kansas \$7 and North Carolina \$6.

This may seem counterintuitive, but Illinois, for example, a state not considered to benefit greatly from the space programme, produces goods and services required indirectly by the recipients of NASA procurement awards: capital goods, electronic components, scientific instruments, chemical products, primary and fabricated metal products, specialized business services and so on. Further, because of the widely based, indirect nature of these economic benefits, Illinois will benefit greatly from NASA procurement spending in other states on a wide variety of programmes, and its benefits are not tied to a specific contract. In this sense, a state such as Illinois is a more certain beneficiary of NASA spending than are some states receiving sizeable prime contract awards.

The data in the table show that, per dollar of direct procurement expenditure, NASA programmes will have widely varying multipliers by industry — these multipliers are interpreted similarly to those for individual states, only here they are computed as the ratio of the total (direct plus indirect) economic benefits to the direct benefits. These range from near six for electronic components, five for electric lighting and wiring equipment and three for chemical

products to near two for motor vehicles and equipment and near one for engines and turbines and aircraft and parts. In other words, the 1987 NASA procurement budget created, indirectly, nearly \$5 in sales in electronic components for every dollar directly spent in that industry, while it created, indirectly, only about 30 cents of sales in the aircraft and parts industry for every direct dollar of expenditure in that industry. The result is not surprising. Aircraft and motor vehicles are final products whose components do not enter into the production of other commodities, whereas electrical and electronic equipment, paints and chemicals are products required in the production of most other goods.

The table also shows the total employment created in various industries. The data here illustrate that the distribution of jobs created by industry differs in important respects from the distribution of sales. Thus, although many jobs are created in industries such as aircraft, ordnance, business services and communications equipment where the generated output requirements are large, employment of equal magnitude is also created in service industries such as wholesale and retail trade, transport, warehousing, restaurants and hotels.

Our findings are significant because we have for the first time estimated the benefits flowing from the second-, third- and fourth rounds of industry purchases generated by NASA procurement expenditures. For some regions these are very important, with ratios of indirect to direct benefits of 8 to 1 and higher. Many workers, industries and regions benefit substantially, and these benefits are much more widespread throughout the United States than has heretofore been realized. We believe our results imply that the economic benefits and costs of space exploration need to be reassessed. □

Roger H. Bezdek and Robert M. Wendling are at Management Information Services Inc., 116 Fourth Street SE, Washington, DC 20003, USA.

1. Morrison, D. C. *Natn. J.* **23**, 1674 (1991).
2. National Aeronautics and Space Administration *Spinoffs* (Washington, DC, 1990).
3. *Economic Impact and Technological Progress of NASA Research and Development Expenditures Report* (National Academy of Public Administration, 1988).
4. US Office of Technology Assessment *Civilian Space Policy and Applications* (Washington, DC, 1982).
5. US Congressional Budget Office *Using Federal R&D to Promote Commercial Innovation* (Washington, DC, 1988).
6. *NASA Annual Procurement Report, Fiscal Year 1987* (Washington, DC, 1988).
7. Management Information Services *The Economic and Employment Benefits to the Nation and to Each State of the U.S. Space Program* (Washington, DC, 1989).
8. Leontief, W. *Input-Output Economics* (Oxford University Press, New York, 1966).

ACKNOWLEDGEMENTS. We thank Jerome Rosenberg for advice. Supported in part by the NASA Alumni League. Complete analyses of industries, states and occupations are available from the authors.

The anthropic view of nucleosynthesis

Our understanding of the Universe may be shaped in part by the circumstance that our knowledge of it derives exclusively from people like ourselves, but what if very different forms of life exist?

THE anthropic principle may be generally known, but that does not mean that it is generally accepted — or even generally understood. And that may be forgivable, for there is a sense in which the anthropic principle does no more than make a virtue of tautology.

In its weakest form, the principle amounts simply to the remark that the only known sources of the observations on which rests our present understanding of the Universe are people like ourselves, members of a species emergent only relatively recently from a long and complicated process of evolution. What this implies is that, of all possible universes, people like ourselves can observe only those old enough to allow for, say, 3,000 million years of natural selection and which also provide the other conditions required for evolution, a reasonable constant ambient temperature for example. So is there much more to say than that the Universe we see is the only one we could see?

The more constructive way to put the argument is to say that because observations of the Universe are collected (so far as is known) only by people, and because the evolution of people requires certain conditions to be met by the physical world, the mere existence of living things makes it possible to say something significant about the Universe as a whole. As will emerge, there is a logical flaw in that element of the anthropic principle, but at least it has the virtue of suggesting that a simple listing of the conditions required for life may have a bearing not only on the question how life began, but on that of how the Universe is constructed as well.

The obvious starting point is the time occupied by biological evolution, which is at least the time elapsed since the earliest forms of microbial life-forms recognizable in Proterozoic rock formations, say 2,700 million years or more. (Those who believe in panspermia — the population of the surface of the Earth by life-forms from space — allow themselves more room for manoeuvre in dealing with the problem of time, but at the cost of straining the attention of their listeners.) As more is learned of the early stages of evolution, and of the physical conditions on the surface of the Earth in early times, it will no doubt be possible to define evermore stringently the conditions that must be satisfied for the emergence of life-forms of any kind, let alone of people, and

thus to say even more about the properties of the whole Universe.

Without mentioning the anthropic principle as such, R. E. Davies and R. H. Koch from the University of Pennsylvania have carried the argument one big step backwards in time by asking what conditions must have been satisfied so that the Solar System should contain the chemical elements required to sustain life (*Phil. Trans. R. Soc.* **334B**, 391–403; 1991). Their starting point is the observation that all life-forms consist not only of carbon, oxygen and hydrogen, but of elements such as phosphorus (as in DNA), iron (as in haemoglobin) and even cobalt (as in vitamin B₁₂). So what would have been required of nucleosynthesis, before the formation of the Sun, so that life could afterwards have emerged as it has done (on one planet of the Solar System?)

The unavoidable starting-point is the assumption that the primordial material consists largely of hydrogen mixed with a tenth as many atoms of helium. All other elements incorporated into living things must have been made at a later stage by nucleosynthesis in supernovae and other kinds of stars. While supernovae are the best known sources of material such as this, it is crucial that some elements — fluorine, for example — are made predominantly on the surfaces of white dwarf stars bound into binary systems in which a larger companion has lost material to the white dwarf.

That conclusion has an immediate bearing on the origin of the material from which the Solar System is made: some of it must have come from earlier supernova explosions and some from the burning of nitrogen nuclei on the surfaces of dwarf stars in binary systems. In other words, the fluorine problem by itself is a proof that the material of which the Solar System is formed came from various stellar sources. That is not surprising. Supernova explosions may produce debris consisting of one or more solar masses of material, but that is necessarily scattered in all directions, so that the formation of a new solar system would require contributions from at least several exploded stars.

The chief value of what Davies and Koch have done may well rest on their compilation of data on the elements involved in living things in a way that must stimulate others to constructive speculation. There is, for example, the observation that even humble *E.coli* depends on

seventeen elements (counting hydrogen, oxygen and carbon), compared with 26 for people. There is also the curious business of why it should be that some eukaryotes concentrate elements such as mercury and cadmium relative to their environment and have then had to evolve detoxifying systems so as to avoid the ill-effects. But with all that said, Davies and Koch are evidently most of all wrapped up with the cosmic significance of their data.

This is how the argument goes. As things are, something like 1.9 per cent of the material of which the Galaxy was formed 20,000 million years ago (or half as long ago, depending on the true value of Hubble's Constant) has been converted by supernova explosions and other means into atoms heavier than helium. Using rough estimates for the efficiency with which supernova explosions effect nucleosynthesis, they conclude that there must have been an average of one supernova explosion every three years since the beginning of the Galaxy and, because the present rate is much smaller, that most of the nucleosynthesis must have happened early in its history. In short, Davies and Koch say, every part of the Galaxy contributed to the formation of the Solar System 5,000 million years ago.

Although the anthropic principle is not mentioned, the bearing of this argument about the origins of the Solar System should be clear: it is not simply that enough time must have passed to allow for the evolution of people, but that that will not happen unless there has been enough nucleosynthesis within some self-contained structure such as a galaxy to make solar systems capable of sustaining life. In short, the only kinds of universes that will be seen by the likes of us must be universes in which there are galaxies whose separate stars are continually refashioning themselves through supernovae.

So what is the logical flaw? Decades ago, a British astronomer of conventional mien was explaining why there could be no other inhabited planet than the Earth to an audience that included the then-youthful Thomas Gold, now to be found at Cornell University. When the speaker had assumed for the umpteenth time that all life is terrestrial, Gold delivered the stage whisper, "Bloody fool, they might look like rocks!"

John Maddox

How chloroquine works

Thomas E. Wellem

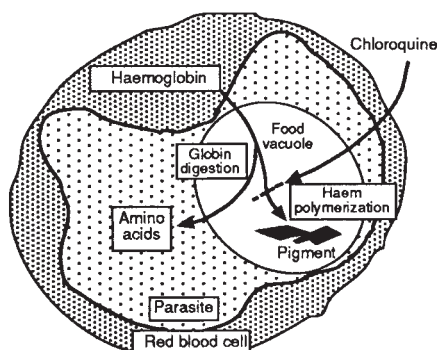
THE antimalarial drug chloroquine has been one of the most successful of antimicrobial agents produced by mankind. Its beneficial effects on public health have been enormous. Yet after nearly a half-century of use, we still do not understand how chloroquine works at the molecular level, or how malaria strains have become resistant to its action. Now, compelling findings reported by Slater and Cerami on page 167 of this issue¹ show that chloroquine interferes with a haem detoxification enzyme that is essential to the survival of malaria parasites in red blood cells.

Malaria is a mosquito-transmitted blood infection caused by *Plasmodium* parasites. The two parasite species responsible for most human infections, *P. falciparum* and *P. vivax*, multiply within and destroy red blood cells in two-day cycles. Rapid propagation of the parasites produces demands for nutrients which are met in part through digestion of the host cell haemoglobin in acid food vacuoles².

But there is a difficulty with the digestion of haemoglobin — its breakdown releases large amounts of a toxic haem moiety, ferriprotoporphyrin IX (FPIX). Intraerythrocytic malaria parasites, unlike their human host, cannot degrade the haem molecules. Instead they sequester FPIX within the food vacuoles as innocuous dark brown granules called malaria pigment or haemozoin (see figure). Physical and chemical studies have shown that the pigment granules are crystals of polymerized FPIX held together by iron-carboxylate bonds between the ferric ions and propionate side chains of adjacent molecules³.

Chloroquine, quinine and related quinoline-ring antimalarial drugs are effective against the intraerythrocytic stages of pigment-producing malaria parasites⁴. Stages in the liver and mosquito, and mature gametocytes that survive without pigment production, are not killed by these drugs. Drug action has therefore been thought to relate to pigment production (8-aminoquinoline drugs such as primaquine act against liver and gametocyte stages by separate mechanisms not considered here). Because chloroquine binds tightly *in vitro* to FPIX in its soluble form, one hypothesis was that chloroquine and related quinoline-ring antimalarial drugs complex with FPIX in the acid food vacuole⁵. Toxic FPIX-drug complexes thus would not be incorporated into pigment, poisoning the acid food vacuole and starving the parasite. But although this hypothesis accounted for the selec-

tive toxicity of quinoline-ring antimalarial drugs against pigment-forming parasites, it left some later experimental results unexplained. Alkalinization of the food vacuole by ammonium chloride, which does not disrupt the FPIX-chloroquine complex *in vitro*, releases accumulated chloroquine from the parasites and prevents toxicity⁶. And affinities of different quinoline-ring antimalarial drugs for FPIX do not correlate with their rank order of antimalarial activity: amodiaquine and halofantrine, for example, are potent antimalarials that do not



THE malaria parasite's haemoglobin digestion pathway and its disruption by chloroquine. Inhibition of haem polymerase blocks the detoxification of ferriprotoporphyrin IX and poisons the food vacuole.

form complexes with FPIX (ref. 7).

Slater and Cerami¹ now reconcile the various experimental findings. A catalytic activity in *P. falciparum* extracts polymerizes haem, presented as FPIX hydroxide or as haemoglobin, into malaria pigment. Chloroquine and six other quinoline-ring compounds (including, significantly, amodiaquine) block haem polymerization according to the extent to which they stem the growth of malaria parasites. These data indicate that quinoline-ring antimalarial drugs act by inhibiting an enzyme that polymerizes and detoxifies FPIX. It seems unlikely that alternative explanations of drug action — for example, that quinoline-ring antimalarial drugs affect lysosome function directly, or that they intercalate into DNA and interrupt gene expression — can account for the facts. Such explanations have to invoke events quite apart from the demonstrated inhibition of FPIX polymerization, and they do not explain either the selective action of quinoline-ring antimalarial drugs against pigment-forming parasites or the relative activities of these drugs against malaria.

For the inhibition of haem polymerase by quinoline-ring antimalarials to be biologically relevant, drug concentra-

tions in the food vacuoles should be close to or more than IC₅₀ values measured *in vitro* (IC₅₀ is the 50 per cent inhibitory concentration, and is 0.12 mM for chloroquine and 0.30 mM for quinine)¹. Red cells infected with malaria parasites can accumulate quinoline-ring antimalarials⁸, and an intravacuolar chloroquine concentration of 1 mM was measured when extracellular chloroquine concentrations were at levels known to inhibit parasite development in culture⁶. Although free-base trapping in the acid compartment was estimated by the Henderson-Hasselbach equation to account for these concentrations⁶ (chloroquine, a diprotic weak base, would partition across membranes with the square of the proton concentration), from other data and comparisons with mammalian cells it seems that parasites have an additional chloroquine-concentrating mechanism⁹, perhaps acting on its protonated form. Also, quinine (a monoprotic weak base) inhibits parasite growth at extracellular levels not expected to produce millimolar concentrations in the acid compartment solely by free-base trapping. So an active concentrating mechanism appears to be necessary to achieve the intravacuolar drug levels that inhibit haem polymerase.

Unfortunately, massive use of chloroquine over the years has resulted in the appearance of chloroquine-resistant *P. falciparum* in nearly every region afflicted by malaria, and resistant *P. vivax* is now known to exist. Resistant *P. falciparum* does not concentrate the drug to the high levels found in sensitive parasites, presumably keeping intravacuolar chloroquine levels below those that would inhibit haem polymerase. An efflux mechanism releases chloroquine from resistant parasites 40 to 50 times faster than from sensitive parasites¹⁰, but a variety of compounds, including verapamil, diltiazem, vinblastine and daunomycin, can partially reverse resistance by inhibiting efflux and raising chloroquine concentrations in resistant

- Slater, A. F. G. & Cerami, A. *Nature* **355**, 167–169 (1992).
- Goldberg, D. E., Slater, A. F., Cerami, A. & Henderson, G. B. *Proc. natn. Acad. Sci. U.S.A.* **87**, 2931 (1990).
- Slater, A. F. G. et al. *Proc. natn. Acad. Sci. U.S.A.* **88**, 325–329 (1991).
- Peters, W. *Chemotherapy and Drug Resistance in Malaria* (Academic, Orlando, 1987).
- Fitch, C. D. in *Malaria and the Red Cell* (eds Evered, D. & Whelan, J.) 222–228 (Pitman, London, 1983).
- Yayon, A., Cabantchik, Z. I. & Ginsburg, H. *Proc. natn. Acad. Sci. U.S.A.* **82**, 2784–2788 (1985).
- Warhurst, D. C. *Ann. trop. Med. Parasitol.* **81**, 65–67 (1987).
- Fitch, C. D. *J. clin. Invest.* **54**, 24–33 (1974).
- Krogstad, D. J., Gluzman, I. Y., Herwaldt, B. L., Schlesinger, P. H. & Wellem, T. E. *Biochem. Pharmacol.* (in the press).
- Krogstad, D. J. et al. *Science* **238**, 1283 (1987).
- Martin, S. K., Oduola, A. M. & Milhous, W. K. *Science* **235**, 899–901 (1987).
- Wellem, T. E. et al. *Nature* **345**, 253–255 (1990).
- Foot, S. J. et al. *Nature* **345**, 255–258 (1990).
- Wellem, T. E., Walker-Jonah, A. & Panton, L. J. *Proc. natn. Acad. Sci. U.S.A.* **88**, 3382–3386 (1991).

strains^{10,11}. These findings make it unlikely that mutations or structural alterations of haem polymerase can account for chloroquine resistance. A mechanism analogous to the multiple drug resistance (*mdr*) mechanism in mammalian tumour cells has been proposed to account for the rapid-efflux resistant phenotype¹¹. Two *P. falciparum* genes with homologies to mammalian *mdr* genes were not, however, linked to chloroquine resistance in a genetic cross, and exceptions to an association of resistance with point mutations in an *mdr*-like gene were also apparent in a population survey^{12,13}. Linkage studies instead place the determinant of the rapid-efflux, chloroquine-resistance mechanism within a segment of *P. falciparum* chromosome 7 (ref. 14).

The loss of chloroquine as a safe,

cheap and effective remedy for malaria already threatens to overwhelm the beleaguered medical services of many countries. Understanding the molecular foundations of chloroquine action and chloroquine resistance — processes that seem to have distinct mechanisms — may point to ways to modify existing quinoline-ring structures, develop therapeutically useful reversal agents or design new classes of haem polymerase inhibitors. In haem polymerization, the parasite is showing us a vulnerable target. The challenge is to find ways to hit it again. □

Thomas E. Wellems is in the Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892, USA.

ANTIGEN PROCESSING

Who needs peptide transporters?

Bernhard Dobberstein

THE discovery of genes in the major histocompatibility complex (MHC) that code for subunits of proteasomes and putative ATP-driven transmembrane transporters has led to a great deal of speculation about their involvement in antigen processing and peptide binding to MHC class I molecules (the developments were discussed at the time in News and Views^{1,2}). The most favoured hypothesis is that proteasomes digest cytosolic proteins into fragments which are then, in an ATP-dependent step, translocated across the membrane of the endoplasmic reticulum (ER) by the transporters. Writing in *Cell*, Lévy *et al.*³ challenge this view. They report that ATP is required for the binding of peptide to MHC class I molecules rather than for peptide transport into the lumen of the ER.

MHC class I molecules are composed of two subunits, a membrane-spanning heavy chain and β_2 -microglobulin, a small exoplasmic protein. Heavy chains are polymorphic and encoded by genes in the MHC. Shortly after their synthesis, heavy chains assemble non-covalently with β_2 -microglobulin in the ER. Binding of peptides, nine or ten amino-acid residues in length, stabilizes the association between heavy chain and β_2 -microglobulin and makes the complex competent for transport to the cell surface, where MHC class I molecules activate cytotoxic T-cells by presenting peptides to them⁴.

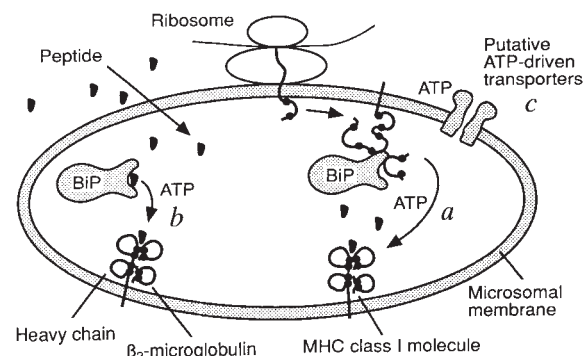
Several mutant cell lines have been characterized in which MHC class I molecules are synthesized but not expressed on the cell surface. Only after the addition of peptides are these molecules found on the cell surface^{4,5}. In

one cell line the defect could be repaired by transfection with a gene encoding a member of the superfamily of ATP-driven transporter proteins⁶. It was therefore proposed that peptide transport from the cytosol into the lumen of the ER is an active, ATP-dependent process. The experiments now described by Lévy *et al.*³ aim at a direct demonstration of the participation of ATP-driven transporters in supplying class I molecules with peptide.

An *in vitro* system was used in which class I heavy chains were synthesized and cotranslationally inserted into membranes derived from the rough ER of lymphoblastoid cells. The membranes, also called rough microsomes, form spherical vesicles that are impermeable to proteases added from the outside. Microsomes have been widely used to study the translocation of nascent secretory proteins. This process requires the presence of an ER signal sequence on the nascent polypeptide, GTP, possibly also ATP, and several cytosolic and membrane factors⁷. Most secretory proteins can be translocated only cotranslationally, but some small ones with relative molecular masses of about 6,000–8,000 (M_r 6–8K) are also translocated post-translationally⁸. Microsomal membranes were generally found to be impermeable to proteins lacking a signal sequence, but small peptides, such as those that are

substrates for an ER oligosaccharyl transferase⁹, can pass through the membrane. When added to cells or microsomes, peptides with 3–5 amino-acid residues are found inside the lumen of the ER or of microsomes, respectively, and become glycosylated. The efficiency of membrane permeation strongly depends on the size and physical properties of the peptides. How these peptides cross the ER membrane is unclear.

To test whether ATP-binding proteins are involved in peptide transport across the ER membrane, Lévy *et al.*³ translated messenger RNA coding for class I heavy chains in a reticulocyte lysate supplemented with rough microsomes from the lymphoblastoma cell line Raji. The microsomes contain endogenous β_2 -microglobulin. After translation, ATP was depleted and a peptide from the nucleoprotein influenza A (NP 384–394) in its biotinylated form was added¹⁰. In the presence of ATP, the peptide stabilized the oligomeric assembly of heavy chains with β_2 -microglobulin and bound to this complex. Depletion of ATP prevented the assembly of heavy chains with β_2 -microglobulin and prevented the binding of peptide. In the absence of ATP, peptide was bound by a chaperone (BiP, the human immunoglobulin-binding protein) and released from it in the presence of ATP; BiP is a soluble heat-shock protein analogue of the ER lumen¹¹, and has been shown to interact with partially folded and misfolded proteins in an ATP-sensitive fashion. The



Proposed ATP-dependent steps in peptide binding to MHC class I molecules. The most favoured view is (a) that ATP is required for folding and assembly of MHC class I heavy chains and β_2 -microglobulin so that peptide can bind. The chaperone BiP is thought to catalyse this process. The alternative view³ is (b) that peptide is concentrated by the binding to BiP and that the ATP is required for the release of the peptide. Whether ATP is also required for membrane translocation (c) of peptides derived from degradation of cytosolic proteins remains an open issue.

two experiments convincingly demonstrate that binding of peptide to MHC class I molecules occurs only in the presence of ATP, and that at least some peptides can enter the lumen of microsomal vesicles in the absence of ATP.

To see whether cytoplasmic domains of transmembrane proteins are involved in peptide transport, Lévy *et al.* incu-

Cost of the charge

SEQUESTERING a charged side-chain in the low-dielectric-constant heart of a globular protein has always been surmised to exact a ruinous thermodynamic cost. Sometimes paired charges have evolved to reduce this penalty. Now S. Dao-pin *et al.* (*Biochemistry* **30**, 11521–11529; 1991) have made two mutants of T4 phage lysozyme, one with lysine for methionine on the inward-facing surface of an α -helix, the other with aspartate for a leucine. The conformations are less stable, but both chains fold, giving products with 35 per cent and 4 per cent activity. The first mutant has been crystallized and the violated helix is appreciably wobbly. The pK of the intruding lysine is 6.5 and the structure survives down to pH 3. Such mutations, then, are evidently not as calamitous as has been supposed.

Old story

CONVENTIONAL thinking has it that clonally reproducing organisms don't last for long in evolutionary terms, at least compared with the sexually reproducing stock from which they stem. Ain't necessarily so, find J. M. Quattro and colleagues in their estimation of the time over which a unisexual, all-female lineage of fish has persisted (*Proceedings of the National Academy of Sciences* **89**, 348–352; 1992). Quattro *et al.* looked at 'post-formatonal' (that is, post-hybridization) mutations in the mitochondrial DNA and allozymes of the unisexual, live-bearing fish *Poeciliopsis monacha-occidentalis*, which occurs in arroyos (streams) in Mexico where the distributions of its two sexual progenitors overlap. As evidenced by mitochondrial DNA restriction-site data, the lineage manifests a high level of genetic diversity and is, the authors estimate, at least 100,000 generations old.

Steps in time

L. R. BRAND, together with Thu Tang, has carried out a further appraisal of how the fossil tetrapod trackways in the Coconino sandstone of northern Arizona came to be formed (*Geology* **19**, 1201–1204; 1991). At issue is whether the sandstone, which is of Permian age, was deposited subaerially, as sand dunes, or under water. Brand and Tang produced diagrams of the orientation of the limb prints and direction of the fossil tracks, and compared them with those made by newts on sand in a tank of flowing water 4 cm deep. In many of the fossil tracks, limb impression and track direction are at different angles, a feature also evident in the impressions left by the living newts as they were drifted by the current. That, say the authors, points towards an underwater origin for at least some of the sandstone.

bated microsomes with protease. This treatment, the authors argue, should have cleaved the cytoplasmically exposed ATP-binding domain of the postulated transporters. Because no effect on peptide binding to MHC class I molecules was evident, they concluded that transporters are not involved in this process. But because the protease sensitivity of the putative transporters is not known, the results are suggestive rather than conclusive.

The idea that ATP-driven transporters are involved in the transport of peptides across the ER membrane arose from analysis of mutant cells defective in surface expression of class I molecules. Mutant T2 cells derived from the B lymphoblastoid cell line T1 have a deletion in the MHC, and they express class I molecules intracellularly but not on the cell surface. Cerundolo *et al.*⁵ found that these cells had lost the ability to present intracellular antigen efficiently, whereas they could present extracellular antigen of eight and twelve amino-acid residues efficiently. One possible explanation for the defect in these cells is that peptides derived from cytosolic proteins do not reach the lumen of the ER. To test this possibility, Lévy *et al.* investigated the ability of microsomes from T1 and T2 cells to take up peptides and assemble them with class I molecules. They found that peptides were taken up into microsomes from T1 and T2 cells with similar efficiencies. But peptide-stimulated assembly of class I molecules was only seen with microsomes of T1 cells, not with those of mutant T2 cells. Lévy *et al.* propose that the defect is in luminal factors needed for peptide binding to the class I molecules.

The authors consider several explanations for the defect in the lumen of microsomes from T2 cells and the ATP requirement for peptide binding to class I molecules. Peptides may become concentrated by binding to BiP and the ATP is needed to release them (see figure). I consider this to be unlikely, because peptide would re-enter the free-peptide pool after release from BiP. Peptides may require trimming in the lumen of the ER before binding to class I molecules. This possibility remains open, however, and needs further experiments with peptides of different size.

Another possibility is that the oligomeric assembly of MHC class I molecules needs a catalyst for proper folding. This is the most likely explanation for an ATP-dependent step, as experimental evidence for a role of ATP for folding and disulphide-bond formation in proteins in the ER will shortly be described by Braakman and colleagues¹². These authors demonstrate that folding of a viral glycoprotein in the ER is an energy-dependent process. In

the absence of ATP proteins misfold into disulphide cross-linked aggregates, but they can be rescued and fold correctly when cells are allowed to regenerate ATP again. Class I heavy chains and β_2 -microglobulin may also require ATP for their folding and oligomerization, and may aggregate in the absence of ATP. Under such conditions peptide binding would not occur.

Is there really no requirement for peptide transporters? Indeed for peptides up to about ten amino acids in length there may be no requirement for active transport across microsomal membranes. But a difference in the permeability of small peptides across membranes of isolated microsomal vesicles from that of the ER membrane in intact cells cannot be ruled out. It will be interesting to see whether peptides applied to the cytosol of cells by either micro-injection or after permeabilization of the plasma membrane can cross the ER membrane in an ATP-independent manner.

The results of Lévy *et al.* raise another important question. What size are the peptides present in the cytosol? In particular, are they the right size to interact with MHC class I molecules? Little is known about the peptide intermediates in the degradation of cytosolic proteins. It is conceivable that proteasomes proteolytically process cytosolic proteins to fragments which can be actively transported across the ER membrane. To draw a more definite conclusion about the requirement for peptide transporters, intact protein substrates, not just trimmed peptides, should be investigated for their requirements for proteolytic processing and membrane translocation. With the *in vitro* system used by Lévy *et al.*, it should be possible to analyse the entire pathway of proteolytic processing of cytosolic proteins, transport of peptides across the ER membrane and binding to the MHC class I molecules. It is likely that several energy-requiring steps and factors mediating charging of MHC class I molecules with peptides still await discovery. □

Bernhard Dobberstein is in the European Molecular Biology Laboratory, Meyerhofstrasse 1, 6900 Heidelberg, Germany.

- Robertson, M. *Nature* **353**, 300–301 (1991).
- Parham, P. *Nature* **348**, 674–675 (1990).
- Lévy, F., Gabathuler, R., Larsson, R. & Kvist, S. *Cell* **67**, 265–274 (1991).
- Neeftjes, J. J., Schumacher, N. M. & Ploegh, H. L. *Current Biol.* **3**, 601–609 (1991).
- Cerundolo, V. *et al.* *Nature* **345**, 449–452 (1990).
- Spies, T. & DeMars, R. *Nature* **351**, 323–324 (1991).
- Gilmore, R. *Current Biol.* **3**, 580–584 (1991).
- Wiech, H., Stuart, R. & Zimmermann, R. *Seminars Cell Biol.* **1**, 55–63 (1990).
- Lau, J. T. Y. *et al.* *J. Biol. Chem.* **258**, 15255–15260 (1983).
- Lévy, F., Larsson, R. & Kvist, S. *J. Cell Biol.* **115**, 959–970 (1991).
- Pelham, R. B. A. *Rev. Cell Biol.* **5**, 1–23 (1989).
- Braakman, I., Helenius, J. & Helenius, A. *Nature* (in the press).

The good companions

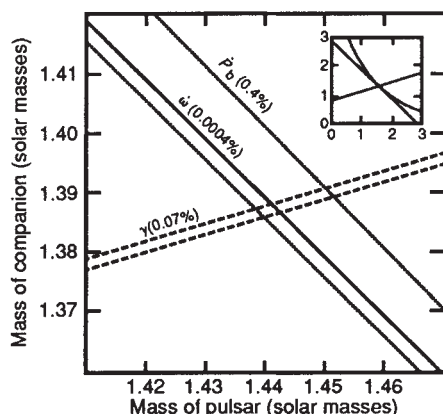
Clifford M. Will

RADIO pulsars that have companions have been extraordinarily useful testing grounds for astrophysics and gravitation. The most celebrated of them is the Hulse–Taylor ‘binary pulsar’ PSR 1913+16, discovered in 1974, which gave the first quantitative evidence for the existence of gravitational waves. Several similar systems, including PSR1534+12 and PSR2127+11C have recently been added to the list. The report by Taylor *et al.* on page 132 of this issue¹ is the latest illustration of just how powerful and sophisticated these systems have become as probes of relativistic gravity. Once again, the results support general relativity. Another report, by Wolszczan and Frail on page 145², announces the detection of a millisecond pulsar blessed with a planetary system. This follows close on the heels of a similar discovery, reported last July³, but this time there are two and possibly three planets.

What makes these discoveries possible is the fact that the pulsars involved rival the world’s ensemble of atomic clocks in timekeeping stability. Evidently they are very old, with magnetic fields and radio emissions weakened by eons of decay, yet with ultra-short rotation periods, in the 1–100 ms range, presumably the result of a more recent spin-up caused by interaction with a companion star involving friction or accretion. Because the magnetic stresses and radiative energy losses are so low, they rotate extremely steadily. Consequently, the slightest displacement relative to the Solar System caused by orbital motion with a companion, even one of planetary size, or the smallest relativistic effect on the apparent pulsar period or on the propagation of the radio signals can be detected through variations in the times of arrival (TOAs) of pulses. In favourable cases, the accuracy in measuring TOAs now approaches the 3 μ s level.

In the case of relativistic binary pulsars, these variations in the TOAs are analysed by fitting the inferred displacements to a two-body relativistic orbit. This fitting yields estimates for five parameters that are associated with the keplerian two-body orbit of ordinary newtonian gravity, such as the orbital period and eccentricity, and a set of parameters associated with relativistic effects. Of these ‘post-keplerian’ parameters the most important ones are the advance of the periastron, $\dot{\omega}$ (the analogue of Mercury’s perihelion shift); the combined effect on the pulsar clock of the gravitational redshift caused by the companion’s gravity field and of the special relativistic time dilation caused

by the pulsar’s orbital motion, denoted by γ ; the rate of change of orbital period $\dot{P}_b$, caused predominantly by in-spiral of the orbit resulting from gravitational radiation damping; and two parameters r and s associated with the delaying effect of curved spacetime around the companion on the propagation of the radio signals out of the binary system. The last effect is the binary-star version of the Shapiro time delay, which has been



Constraints on masses of pulsar and companion from data on PSR1913+16, assuming general relativity to be valid. Width of each strip in the plane reflects observational accuracy, shown as a percentage. Inset shows the three constraints on the full mass plane; the intersection region has been magnified 400 times for the main figure.

tested to 0.1 per cent in the Solar System using radar tracking of planets and satellites⁴.

Two things make binary pulsars unique tools for testing general relativity: the emission of gravitational radiation and the presence of strong-field effects. In the Solar System, gravity is weak everywhere: the newtonian gravitational potential never exceeds $10^{-5} c^2$, where c is the speed of light. As a result, Solar System experiments probe only the weak-field, slow-motion corner of the space of predictions of gravitation theory. Also, the effects of gravitational waves in the Solar System are negligible.

Neutron stars, by contrast, are strong-field objects. In their interiors, the gravitational potential can be as high as $0.2 c^2$ (indeed, even invoking newtonian language here is a bad approximation). Ironically, this makes not one whit of difference as far as general relativity is concerned. When tidal effects can be ignored, as is usually the case in binary pulsars, the motion of a body and the gravitational waves produced by its motion are independent of its internal structure. In such situations, for a body of given mass, it matters not whether it

is a normal star, a neutron star or a black hole. This ‘effacement’ of the internal structure is called the strong equivalence principle (SEP), a property of general relativity. Most other theories violate the principle at some level.

One (oversimplified) way, to think about such non-SEP effects in binary pulsars is to say that a neutron star can be characterized by different masses, each differing from the other by amounts dependent upon the mass-energy associated with the internal strong gravitational fields. One of these might be the inertial mass, another might be the mass that appears in Kepler’s third law for the orbital period, a third mass might appear in the periastron advance formula, and so on. In general relativity, all these masses are identical. Measuring the various keplerian and post-keplerian parameters constrains these masses. The extent to which these constraints all lead to a single, unique value for the mass of each member of the binary system is a measure of the validity of general relativity. The figure shows how the $\dot{\omega}$, γ and $\dot{P}_b$ constraints from PSR1913+16 overlap successfully, assuming general relativity. The results give 1.4411 and 1.3873 solar masses for the pulsar and its companion, respectively, each accurate to 0.05 per cent. The measured value of $\dot{P}_b$ is consistent with general relativistic gravitational-wave damping at the 0.5 per cent level.

More remarkable in some ways are the results reported by Taylor *et al.* for the recently discovered⁵ binary pulsar PSR1534+12. Here the data span of about a year is not yet sufficient to give a meaningful value for $\dot{P}_b$, although it is expected to do so eventually. Instead, PSR1534+12 yields interesting values for the ‘time delay’ parameters r and s . The Hulse–Taylor pulsar tells us little about these parameters as its modest orbital inclination relative to the plane of the sky means that the pulsar signals do not pass that close to the companion, with the result that the Shapiro time delay is suppressed. By contrast, PSR1534+12 is seen to be nearly edge on, so once in each orbit the pulsar signal encounters the companion’s strong local spacetime curvature on its way to the Earth, and thus suffers a significant Shapiro delay. By combining the values of $\dot{\omega}$, γ , r and s , Taylor *et al.* come up with a test of strong-field gravity in PSR1534+12 that does not rely upon radiative effects.

One problem is finding an interesting theory to test. Metric theory alternatives to general relativity are of roughly two types: those that are ‘simply connected’ to general relativity and those that are not. Simply connected means a theory whose difference from general relativity is measured by one or more continuous parameters, such that in an appropriate

limit of the parameters, the theory tends smoothly to general relativity in all its predictions. The 1960s vintage scalar-tensor theory of Brans and Dicke (which was based on earlier theories by Jordan and Fierz) is the classic example. There a parameter ω_{BD} is an inverse measure of the strength of a scalar field added to the usual spacetime metric tensor field. As ω_{BD} tends to infinity, Brans-Dicke theory tends to general relativity.

Non-simply connected theories are so qualitatively different from general relativity in their formulations, that they do not formally tend to it in any limit. Nevertheless, many theories in this class can be made to agree with general relativity in the weak-field, slow-motion regime of the Solar System.

Binary pulsar tests have proven to be deadly for non-simply connected theories, because their qualitative divergences from general relativity usually show up with a vengeance in strong-field and radiative situations. The best example of this was the 'bimetric' theory of Nathan Rosen, which agreed with Solar System tests but failed spectacularly in the binary pulsar⁶.

The same cannot be said for simply connected theories. Not that they do not deviate from general relativity in binary pulsar systems, for they do, but they also generally deviate from it in the Solar System, and so are already constrained, usually more strongly than is currently achievable in binary pulsars. Current Solar System tests of light deflection and Shapiro time delay (both around 0.1 per cent) constrain the Brans-Dicke theory's ω_{BD} to be greater than 500 (ref. 4). This makes the theory so close to general relativity (roughly within factors of $1/\omega_{BD}$ in all its predictions) that it easily satisfies all the binary pulsar constraints⁷. Put differently, the results quoted for PSR1913+16 by Taylor *et al.*¹ provide the constraint $\omega_{BD} > 100$, which is not competitive with the Solar System bound.

This raises a theoretical question. Is it possible to find a simply connected theory of gravity that is not constrained by Solar System tests, but that can be constrained by binary pulsar tests? Damour and Esposito-Farese have shown that it is (T. Damour and G. Esposito-Farese *Class. Quant. Grav.*, manuscript submitted). They cook up a scalar-tensor theory *à la* Brans-Dicke in which two scalar fields appear, tuned so that their effects in the weak-field limit of Solar System tests cancel, making the theory indistinguishable from general relativity for that application. In the strong-field and radiative limits, however, the scalar fields combine to produce observable differences. Yet as the theory's two parameters β' and β'' tend to zero, the theory does tend smoothly

to general relativity. As Taylor *et al.* report, the two binary pulsar systems together provide a strong constraint on these parameters, consistent with zero.

The Wolszczan-Frail planetary system² will not have much impact on general relativity, because the planets are too light to emit significant gravitational radiation, and their orbits are too circular to reveal a meaningful periastron. But it does present a challenge to astrophysicists to explain how and when such planets formed, and how they survived the various cataclysms that usually afflict the lives of millisecond pulsars. It also suggests that planetary systems in the Galaxy may be more common than has hitherto been evident from optical

NEURAL NETWORKS

Learning from your neighbour

Graeme Mitchison and Richard Durbin

WHAT can artificial neural networks tell us about the brain? One view is that they can be used to explore the consequences of different synaptic learning rules in a simplified formal setting. However, the most powerful learning algorithms, such as back propagation¹, need an external 'supervisor' to correct the mistakes made by the network, which is an unrealistic requirement especially for early stages of sensory processing. How, therefore, does one learn effectively without a supervisor? On page 161 of this issue², Becker and Hinton propose an answer to this question. Theirs is not the first unsupervised learning algorithm, but they take a new approach which has a paradoxical charm: in effect, different pieces of the inputs train each other.

The goal of an unsupervised learning algorithm is to extract meaningful features or variables from a set of input patterns. For example, we can try to find those features that allow the data to be reconstructed as faithfully as possible. This is the goal of principal component analysis, a standard tool of engineering and statistics. By identifying the combinations of inputs with maximum variance, it finds the variables that can be most effectively used to characterize the inputs. Remarkably enough it turns out that the first neurobiological learning rule to be formulated, Hebb's rule³, is closely related to principal component analysis. Given a simple neural-network model consisting of a single unit, Hebb's rule results in that unit extracting the largest principal component, assuming some form of normalization of synaptic connection strengths^{4,5}. With a small amount of modification, a set of units can be made to learn not just the largest component, but a set of components which together capture the greatest

searches (which have found nothing). On the other hand, I wouldn't lay odds on the existence of life on *these* planets. □

Clifford M. Will is in the McDonnell Center for the Space Sciences, Department of Physics, Washington University, St Louis, Missouri 63130, USA.

1. Taylor, J. H., Wolszczan, A., Damour, T. & Weisberg, J. M. *Nature* **355**, 132-136 (1992).
2. Wolszczan, A. & Frail, D. A. *Nature* **355**, 145-147 (1992).
3. Bailes, M., Lyne, A. G. & Shemar, S. L. *Nature* **352**, 311-313 (1991).
4. Will, C. M. *Science* **250**, 770-776 (1990).
5. Wolszczan, A. *Nature* **350**, 688-690 (1991).
6. Will, C. M. & Eardley, D. M. *Astrophys. J. Lett.* **212**, L91-L94 (1977).
7. Will, C. M. & Zaglauer, H. W. *Astrophys. J.* **346**, 366-377 (1989).

part of the variance^{6,7}.

Principal components appear in at least some cases to be involved in biological processes. In the local processing of visual images, for example, the principal components include edge segments, which are among the first features extracted in primary visual cortex⁸. However, other important variables, such as stereoscopic disparity, will not be explicitly extracted. Becker and Hinton show how one could set about extracting these more elusive variables. One way to describe their approach is that they assume that the interesting properties are more stable than the noise. For example, the depth of a surface, as measured by stereoscopic disparity, will tend to vary smoothly in scanning across an image, whereas the local pixel intensities may vary rapidly because of texture.

Consider a system looking at two neighbouring, non-overlapping patches, and suppose that, corresponding to each patch, there is a unit whose inputs come from that patch only. One could try to make the units extract a stable property by requiring that they both perform the same computation on their input and by minimizing the difference in their responses. But then they might end up both doing nothing (that is, give a zero response). To avoid this, one could try to mimic hebbian principal component learning, and ask the units to maximize the variance in their responses. Becker and Hinton combine these requirements by making the units maximize the variance of the sum of their outputs divided by the variance of their difference.

On the assumption that both the underlying variable and the noise have a gaussian distribution, this is equivalent to maximizing the mutual information of the two outputs. Here one can see parti-

cularly clearly how the algorithm works: the mutual information can be large only if, first, the units convey information (that is, they behave nontrivially) and if, second, they respond similarly, so they share this information. The Hebb rule essentially imposes the first constraint alone. By adding the second constraint the new rule allows information to be thrown away when it is not shared by other patches.

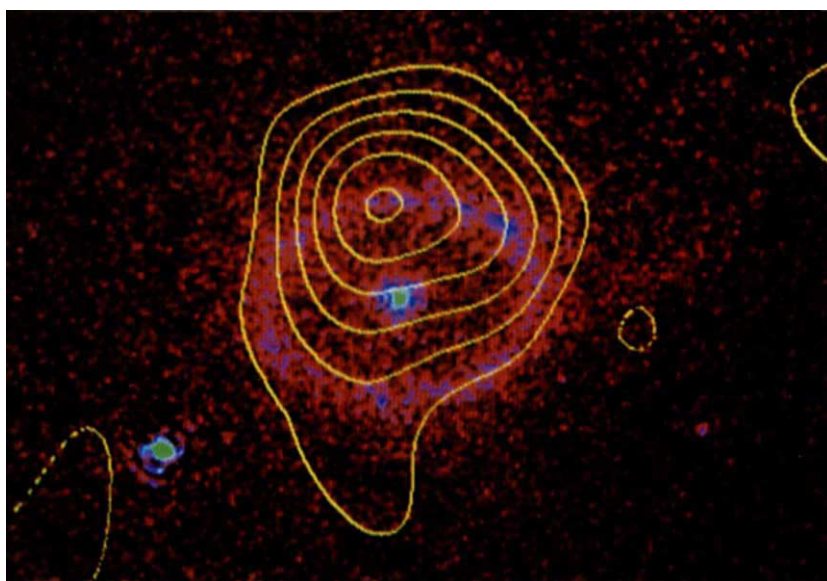
Maximizing mutual information can also be interpreted as prediction, because each unit can be used to predict the behaviour of neighbouring patches. The notion of prediction is more general than that of stability: we can look for properties that predict future inputs, or predict one set of sensory data through another sensory modality. Prediction can help to complete or interpret missing data, and where prediction fails something interesting is likely to be happening. For example, places where disparity changes sharply will usually correspond to the edges of objects.

Neural networks are inspired by real neurons, but is there any reverse flow of inspiration? Might a rule such as this operate in the brain? It seems unlikely that neurons compute something as mathematically complex as the ratio of variances, let alone the determinants which occur in the more general expression for more than two units. Furthermore, some of the difficulties of back propagation apply to the multilayer version of this algorithm, which must somehow feed back a complex error signal to earlier stages in the neural pathway. But it is important not to be too intimidated by the mathematical formulation. After all, principal component analysis, which in its standard form requires matrix inversion, might seem an unlikely operation for neurons to accomplish. Yet it can be carried out by suitably organized hebbian machinery. It seems likely, in fact, that there are natural ways for neurons to carry out Becker and Hinton's kind of analysis, or something very close to it, and this may provide another clue to help us explore synaptic learning rules in the brain. □

Graeme Mitchison is in the Physiological Laboratory, University of Cambridge, Cambridge CB2 3EG, UK. Richard Durbin is in the MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

1. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Nature* **323**, 533–536 (1986).
2. Becker, S. & Hinton, G. E. *Nature* **355**, 161–163 (1992).
3. Hebb, D. O. *The Organization of Behavior* (Wiley, New York, 1949).
4. Oja, E. *J. math. Biol.* **15**, 267–273 (1982).
5. Linsker, R. *Computer* 105–117 (March 1988).
6. Sanger, T. D. *Neural Networks* **2**, 459–473 (1989).
7. Földiák, P. in *Proc. Int. Joint. Conf. Neural Networks* Vol. 1, 401–405 (IEEE, New York, 1989).
8. Hubel, D. H. & Wiesel, T. N. *J. Physiol.* **160**, 106–154 (1962).

Radio days of a remnant supernova



In the history of supernova 1987A, now almost five years old, radioastronomers have so far had a negligible role. Apart from a brief initial outburst of radio emission, lasting no more than a few days, the expanding nebula set into motion by the explosion has been quite invisible at radio frequencies, and the steady thinning and cooling of the ejected material, and its interaction with the circumstellar material that surrounded the progenitor, has been followed largely through ultraviolet, optical and infrared observations. But elsewhere in this issue (*Nature* **355**, 147–149; 1992), L. Staveley-Smith *et al.* describe their detection of radio emission from the remnant, illustrated here overlaid on an optical picture from the Hubble Space Telescope. Evolution of the radio remnant over the coming years will provide a new tool to dissect the progress of the expanding remnant.

The key to understanding the remnant of SN1987A lies in the nature of its unusual progenitor star, which was first a red-giant, then a blue giant, before it exploded. In its red-giant phase, the star threw off a dense, slow-moving wind, which was succeeded by a more tenuous but faster wind from the blue-giant. The circumstellar material of the progenitor at the moment of explosion therefore consisted of a hot thin gas cocooned inside a cooler, thicker shell, with a shock wave created at the boundary as the blue-giant wind ran into the red-giant wind.

The first brief flash of radio emission, reported by A. J. Turtle *et al.* (*Nature* **327**, 38–40; 1987), was a very minor part of the initial supernova outburst, and was probably attributable to the propagation of the shock wave from the explosion through the thin material immediately surrounding what had been

the progenitor star. According to R. A. Chevalier (*Nature*, in the press), the emission now detected by Staveley-Smith and colleagues is due to the same expanding shock finally reaching the outer edges of the old blue-giant wind, just before it runs into the denser red-giant wind. Chevalier predicts that as the expanding ejecta passes through this boundary layer, the radio signal will rise and then diminish again, a signature which should be seen sometime during 1992.

After this transient appearance, SN1987A is unlikely to emerge as a fully formed radio supernova remnant for some time. The ages of radio remnants seen in other galaxies as well as our own are typically measured in hundreds of years at least, and there have been few opportunities for astronomers to observe a supernova at close enough hand to see the radio remnant arise from the expanding nebula. Before the advent of SN1987A, astronomers had to make do with studies of supernovae in other galaxies, and those that are detectable at radio frequencies have been either mature remnants or very new ones, which have faded within a few years.

Just before the radio recapture of SN1987A, however, J. Cowan, *et al.* (*Astrophys. J.* **379**, L49–L51; 1991) spotted the reappearance of a 20-year old supernova, SN1970G in the galaxy M101, that had been radio-bright for about three years after outburst but which had then sunk below detectability. It is thought that the progenitor of SN1970G was, like that of SN1987A, a fairly massive star, and the explanation for the reappearance of radio emission from the former after 20 years and from the latter after five may be essentially the same.

David Lindley

Building blocks on shaky ground

Arch C. Johnston

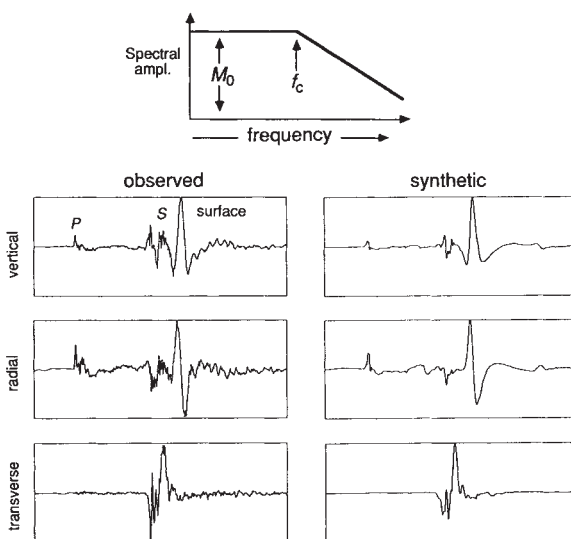
ALL sciences are anchored on a few fundamental building blocks; in seismology there are none more fundamental than that earthquakes are shear dislocations on surfaces known as faults. A much more recent block was added in the 1960s with the notion of self similarity¹; that is, aside from a size scale factor, all earthquakes are alike in the delicate balance between the size of faults, the amount of slip occurring on them, and the drop in stress between pre- and post-rupture states. A new study², by Hough and Seeber of the Lamont-Doherty Geological Observatory, of a modest quake in a most unlikely location just north of New York City has reopened a simmering debate over just how similar earthquakes really are.

The analysis of Hough and Seeber seemingly reaffirms self similarity on one front but calls it into question on another. Good-quality seismic data were obtained from a body-wave magnitude $m_b = 4.0$ earthquake (comprising two sub-events) and its aftershocks occurring near Ardsley, New York, north of Manhattan, in October 1985. These data lend support to the notion of self similarity in terms of size. For, by the authors' interpretation and within the resolution of their instruments, the stress drop following each sub-event or aftershock showed no significant dependence on the strength of the shock.

On the other hand, the average stress drop for the whole sequence was 200 bars, and the highest value measured (one of the main subevents) was more than 500 bars, significantly exceeding the worldwide average for shallow crustal earthquakes, which falls somewhere in the range 10–100 bars. This finding supports the controversial contention³ that earthquakes in stable plate interiors relieve more stress, by about a factor of six, than those at plate margins and in active tectonic zones, even though there is no reason to expect any gross variation in the rheology of crustal rocks. If true, this geographical difference in stress release would itself constitute a violation of self similarity, implying that the less active midplate faults are stronger and harder to break.

Questions about stress drop and self similarity have survived unresolved because extremely high-quality data are required to answer them. One rarely

finds the happy combination of an earthquake sufficiently large to yield a good signal-to-noise ratio and modern high-dynamic-range instrumentation that remains on scale for such an event close to the epicentre, particularly in earthquake-poor plate interiors. The Ardsley quake was recorded by the Lamont-Doherty team at Palisades, only 8 km away. By a stroke of luck the azimuth to the station was in a low-amplitude nodal direction of the primary (P) wave radiation



Top, an idealized earthquake spectrum, illustrating the low frequency zero-slope spectral level that is proportional to earthquake size (seismic moment, M_0) and the corner frequency f_c , above which the spectrum falls off. Below, synthetic seismogram comparison with actual data from the broad-band seismic station at Cathedral Cave, Missouri, 210 km distance from the 4 May 1991, $m_b = 4.6$ New Madrid zone earthquake. All wave groups (P, S and surface) are closely matched only if a seismic moment $M_0 = 1.9 \times 10^{22}$ dyne cm, a hypocentral depth of 7–8 km and a left-lateral strike-slip faulting mechanism are used. (Courtesy, R. Herrmann.)

pattern and, except for a few minor clip points, the initial P-wave signal remained on scale.

With an on-scale seismogram, the powerful techniques of Fourier spectral analysis can be brought to bear to obtain two crucial measures of the source earthquake: its seismic moment (M_0) and corner frequency (f_c). The first represents the strength of the quake in terms of the force couple that is equivalent to the shear dislocation; the second leads to direct estimates of both fault dimension and stress drop. The seismogram is digitized and Fourier transformed from the time domain to the frequency domain, in which its low-frequency spectral level is proportional to M_0 , and f_c is the point at which this zero-slope level abruptly decreases, usually as the inverse square of frequency (ref. 4; see figure).

The determination of seismic moment has become much more reliable in recent years, with the advent of synthetic seismogram modelling techniques, in which M_0 simply becomes a scale factor to match observed to synthetic waveforms. Robust determination of corner frequency, however, is fraught with difficulty, because attenuation of the seismic signal along the ray path through the crust introduces a curvature in the spectrum just where the observation of a sharp f_c is critical. Moreover, stress drop is proportional to the cube of f_c , so that slight uncertainties in f_c yield much larger uncertainties in stress drop, and it is small wonder that seismologists generally view with healthy scepticism any quoted stress-drop values (except their own).

Even with the favourable conditions of the Ardsley earthquake study, the data, although sufficiently good to advance the debate about self similarity, are not of the necessary quality to resolve it. For example, the dynamic range of the Palisades instrument was too low; only the first 0.7 seconds of the primary wave was close to being on scale. Seismic moments could not be directly computed because of uncertainties in seismometer response, so they were estimated from m_b . And even with a station within 8 km, attenuation remains a source of considerable uncertainty: the spectra of aftershocks from Palisades could be modelled equally well with corner frequencies varying by a factor of two by adjusting the attenuation factor. Because of these uncertainties, one could argue that the Ardsley sequence *does* show a dependence of stress drop on magnitude, just the opposite conclusion to that of Hough and Seeber.

Data from recently deployed seismic instruments in the New Madrid seismic zone of the central United States illustrate the excellent results that are possible when the stringent instrumentation requirements for quality earthquake source studies are indeed met. On 4 May 1991, the largest New Madrid earthquake in 15 years ($m_b = 4.6$) occurred near Risco, Missouri. Three different types of instruments captured this event. A new broadband, three-component IRIS (Incorporated Research Institutions in Seismology) station nearby provided an excellent on-scale record from which a synthetic seismogram match by R. Herrmann of St Louis University (see figure) provided a tightly constrained $M_0 = 1.9 \times 10^{22}$ dyne cm. The PANDA (Portable Array for Numerical Data Acquisition) instruments densely deployed in the region by Memphis State

University pinpointed the earthquake's depth and mechanism of faulting⁷, parameters essential to producing synthetic waveforms. Finally a strong-ground-motion instrument operated by NCEER (National Center for Earthquake Engineering Research) was sited only 3 km away. Spectra of its data⁸, with attenuation minimized by the short ray path, yield a well-constrained f_c of 2.8–3.2 hertz. Combining the results from these very different but complementary types of instruments allows one to derive a stress drop for the Risco earthquake of 103 ± 20 bars. This one data point does not answer the self-similarity controversy, but another dozen or so such well-determined points will.

The New York study showed what is possible with a little luck and squeezing the last drop of analysis out of data with quality limitations. The New Madrid study shows what is possible with GEOMETRY

new generation state-of-the-art instruments . . . and also a bit of luck. The good news is that these are significant advances towards understanding the source processes of midplate earthquakes. The bad news is that economic realities will severely limit deployment of the advanced instrumentation to but a fraction of that needed. □

Arch C. Johnston is at the Center for Earthquake Research (CERI), Memphis State University, Memphis, Tennessee 38152, USA.

1. Aki, K. *J. geophys. Res.* **72**, 1217–1231 (1967).
2. Hough, S. E. & Seeber, L. *J. geophys. Res.* **96**, 18183–18185 (1991).
3. Scholz, C. H., Aviles, C. A. & Wesnousky, S. G. *Bull. seism. Soc. Am.* **76**, 65–70 (1986).
4. Brune, J. N. *J. geophys. Res.* **75**, 4997–5009 (1970).
5. Chiu, S. C. Pap. presented at 63rd A. Meet. Eastern Sect. seism. Soc. Am., Memphis, 14–16 Oct. 1991.
6. Friberg, P. A., Jacob, K. H. & Hough, S. E. Pap. presented at 63rd A. Meet. Eastern Sect. seism. Soc. Am., Memphis, 14–16 Oct. 1991.

Another harmony of the spheres

Nelson Max

IN 1611, Kepler¹ described the face-centred cubic arrangement (Fig. 1), and claimed that it was the densest way of packing equal spheres in three dimensions. This claim defied proof until W. Hsiang tackled it last year².

Face-centred cubic (f.c.c.) packing is used by greengrocers to display piled oranges: the bottom layer is a square array, and succeeding layers are placed in the holes between any four spheres in the layer below. Physicists and chemists have long accepted that there is no denser packing for hard spheres, but mathematicians have been unable to prove it. In fact, David Hilbert³ included the proof in number 18 of his 1901 list of unsolved problems in mathematics.

In 1831, Gauss⁴ proved that f.c.c. is the densest of all *lattice* packings — in which the spheres are all arranged at regular intervals along the three lattice directions. But this left open the possibility of denser packings, either regular nonlattice packings like the hexagonal close packing, whose unit cell contains two spheres, or irregular patterns like quasi-crystals, which never repeat in any direction.

Wu-Yi Hsiang became interested in the problem while preparing to teach a course in classical geometry. Two preprints⁵, totalling 150 pages, give many of the details of his proof. Remarkably, he uses no modern mathematics, only high-school algebra, vector identities, a little calculus and a lot of spherical geometry and spherical trigonometry. Hsiang partly attributes his success to the training he received in this

'old-fashioned' mathematics as a student in Taiwan.

The basic idea in the density calculation goes back to Kepler¹, who observed that as the fruity parts of a pomegranate expand against each other, they are mostly squeezed into the shape of the rhombic dodecahedron (Fig. 2). In the present case, one can imagine the spheres expanding radially like balloons, stopping locally where they collide.

To make this precise, one can define the local cell for a sphere in a packing as the set of all points which are at least as close to the sphere's centre as to any other sphere centre. (This is also called a Voronoi cell, or Dirichlet domain.)

Hsiang defines the local density for a sphere in a given packing as the ratio of the volume of the sphere to the volume of its local cell. Because the local cells fill up the total volume into which the spheres are packed, the global density (the ratio of the volume of all the spheres to this total volume) is equal to the ratio of the volume of one sphere to the average of the volumes of the local cells.

In the plane, the relationship of maximum local density to maximum global density is simple. Hsiang⁶ takes two pages to prove that the regular hexagon has the minimal area of a local cell in any packing of

circles. As regular hexagons can form a lattice packing with all the local cells identical, the maximum local density is also the maximum global density. But in three dimensions, the minimum-volume local cell does not occur in a lattice packing, making the proof much more difficult.

The local cell for a sphere in f.c.c. packing is a rhombic dodecahedron, bounded by the tangent planes where it touches its 12 closest neighbours (Fig. 2). If each sphere has radius 1, its volume is $\frac{4}{3}\pi$ and the rhombic dodecahedron has volume $4\sqrt{2}$, so the local density, the ratio of these, is 0.74048. And because f.c.c. packing is a lattice packing, this local density is also the global density.

On the other hand, the local cell of smallest volume is a regular (pentagonal) dodecahedron, arising when 12 neighbouring spheres touch a central one at the centres of the 12 pentagonal faces. This regular dodecahedron has smaller volume than the f.c.c. cell, so the local density is higher. However, regular dodecahedra cannot be arranged into a lattice packing, so this does not imply a high global density. Indeed, this arrangement of 12 neighbours causes a compensating increase in their own local cells. Because the 12 neighbouring spheres do not touch each other, their ballooning local cells expand across the gaps between them. Hsiang managed to quantify this extra volume, and show that the total volume of the cluster of 13 local cells is always at least as large as in a cluster of 13 f.c.c. local cells. The upper bound this implies for the global density of any packing then follows from an averaging argument.

To prove this lower bound on the volume of a cluster of 13 local cells, Hsiang had to consider a second layer of spheres around the first layer of 12, so as

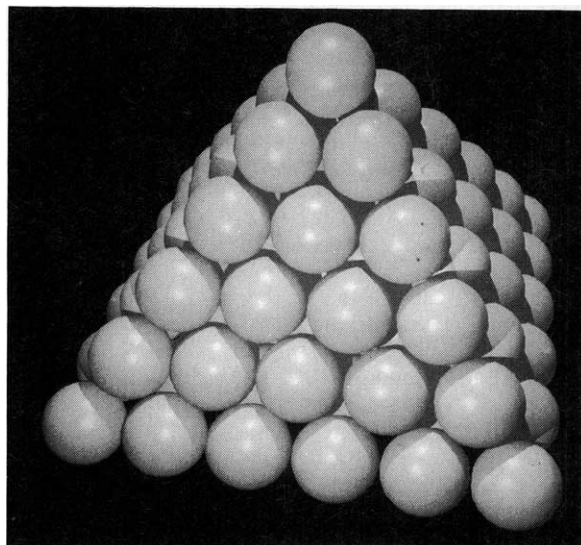


FIG. 1 Face-centred cubic packing, familiar from fruit stalls throughout the world.

to enclose their local cells. In the f.c.c. packing, the second layer has 42 spheres. In the related hexagonal close packing, it has 44, giving a total of 57 spheres when the inner 13 are included. Thus up to $57 \times 3 - 6 = 165$ parameters are needed to specify all configurations of two layers, making their classification and analysis a formidable task.

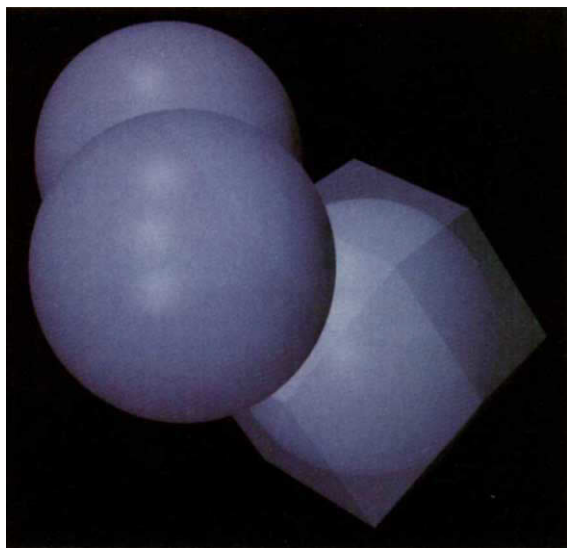


FIG. 2 The local cell for a sphere in f.c.c. packing. The 12 surrounding spheres are centred on the midpoints of the sides of the cube of side $2\sqrt{2}$ (two are shown).

Hsiang's strategy was to eliminate most configurations as giving much too large a volume, and then to show that the remaining possibilities are close to one of a few key cases, like those mentioned above, which are analysed more carefully. It is this classification that takes up most of the second, longer preprint⁵, and it took Hsiang over a year to organize. Although he gives some explicit constants to separate the cases, many distinctions in the current draft are described qualitatively. This makes for easier reading but harder verification. Hsiang decided that to include the full details would make a much longer and more boring paper.

What is the current status of the proof? The original announcement, submitted to the *Bulletin of the American Mathematical Society* late in 1990, was rejected because the detailed proof had not yet been written down, but now it has. The draft will be circulated for comments to interested mathematicians for a few months. Hsiang is also presenting the details to the students in his graduate mathematics class. The current draft is the second, and it will probably be revised again before submission for publication. I have already pointed out a couple of minor, easily corrected misstatements, and presumably others will suggest improvements. But I doubt anyone will have the patience to repeat Hsiang's year-long verification that all

possibilities are covered.

Why not do the detailed verification by computer, as in the four-colour map theorem? This does not seem easy. The four-colour problem was basically a combinatorial one, whereas the packing problem involves a large number of real position variables and a careful choice of technical parameters to distinguish the cases. Hsiang did not use a computer at all, although one sits in his office. He did the many spherical trigonometry and volume calculations on his trusty TI 35+ pocket calculator.

And the future? Hsiang is already working on the four-dimensional case. There is a well-known lattice packing called D_4 , in which each sphere is surrounded by 24 touching neighbours. Hsiang claims to have a proof that this is the only configuration in which 24 spheres can touch a central one. He conjectures that this D_4 packing will have the minimum local cell volume. This would give it the maximum density, as in the two-dimensional case, because it is a lattice packing.

As for applications to the physical sciences, Hugh DeWitt at the Lawrence Livermore National Laboratory has written to Hsiang to recommend "a very similar problem in three dimensional space that is important in physics right now — the lowest possible energy of point charges confined in a finite volume, and with a uniform fixed background of opposite charge to give overall charge neutrality. This system is called the One Component Plasma. It has very important applications as a model for the interior of white dwarf stars and to some terrestrial systems as well . . . the b.c.c. [body-centred cubic] lattice is very slightly lower in energy than the f.c.c. lattice in contrast to the hard sphere problem. As with the hard sphere packing problem, all physicists are convinced that the b.c.c. lattice is the lowest energy state, but nobody knows how to prove it". □

Nelson Max is in the Department of Applied Science, University of California, Davis, California 95616, USA.

1. Kepler, J. *On the Six Cornered Snowflake* (Oxford University Press, 1966).
2. Hsiang, W. *Bull. Brazilian math. Soc.* (in the press).
3. Hilbert, D. *Archiv. Math. Phys.* **1**, 44–63 (1901).
4. Gauss, C. *Werke II* 188–196 (1876).
5. Hsiang, W. Reports PAM-530; PAM-535 (Center for Pure and Applied Mathematics, University of California, Berkeley, 1991).
6. Hsiang, W. Report PAM-536 (Center for Pure and Applied Mathematics, University of California, 1991).

Artesian diver

INDUSTRIAL society produces an ever-increasing quantity of rubbish. Burning it in air, even in specialized furnaces, gives rise to air pollution. Daedalus now proposes to burn it in water instead.

This process is already used to destroy some specialized chemical wastes. They are dispersed in oxygen-bearing water which is heated to criticality at 380°C and 220 atmospheres pressure. Supercritical fluids are powerful solvents and reaction media. The organic material dissolves and oxidizes flamelessly and completely; the products remain safely in solution. To scale this process up would stretch high-pressure engineering to its limits. But Daedalus is undaunted. He suggests sinking an artesian well 2.2 km deep, when the hydrostatic pressure at the bottom will reach 220 atmospheres.

The necessary temperature of 380°C needs a bit more cunning. Daedalus would like to use deep geothermal heat, thus achieving both pressure and temperature naturally; but in most places geology is against him. Instead, he recalls a scheme for keeping a diver warm in the icy ocean by beaming microwaves down his air pipe, which is configured as a wave guide. So by pumping air down an artesian well through a pipe which also acts as a wave guide, copious microwave heating could be provided at the bottom. The water there should rapidly reach supercriticality. Even better, microwaves promote chemical reactions wonderfully. Many traditional laboratory processes are greatly accelerated by microwave heating. Oxygenated, supercritical, microwave-heated water should dissolve and burn all forms of waste material rapidly and completely.

So Daedalus's artesian waste-combustor consists simply of two artesian wells connected at the bottom. Air, microwaves and every kind of rubbish and sewage sludge will be pumped down one of them; steam, carbon dioxide and hot water will fizz violently out of the other, providing copious heat and power. Even allowing for the power fed back to generate the microwaves, the installation will make a splendid energetic profit.

Under the ferocious conditions at the bottom, combustion will be guaranteed complete. All organic material will go to carbon dioxide, steam and nitrogen; minerals will be decomposed and equilibrated; the chlorine in polychlorinated biphenyls and PVC will be perfectly disengaged with no parts per billion of dioxin or chloroform to worry environmentalists. The outflow water will be utterly sterile, subtly mineralized and pleasantly carbonated. If its origin can be kept quiet, it should outsell the smartest mineral waters.

David Jones

Crystallization by centrifugation C₆₀ separation on coal

SIR — Dover¹ pointed out that “those investigating protein structure by X-ray crystallography depend on a reliable supply of sufficiently large, well-ordered single crystals on which to work”. The production of good crystals forms a critical rate-limiting step in undertaking three-dimensional X-ray analysis. Methods for the crystallization of proteins have been intensively studied². The requirement for new approaches to the problem has led to experiments at zero gravity in space, high *g*-force and pressures²⁻⁴. At normal pressure and gravity, proteins have traditionally been crystallized by inducing supersaturation using parameters including the addition of precipitant, for example salts or alcohols, altering the pH towards the isoelectric point, slowly changing the temperature of the system or concentration by evaporation. These crystallization experiments are performed over a period of time which can be days, weeks, months or even years. I have devised a new method for rapid crystallization by centrifugation.

I have grown large single crystals of a newly purified aspartic proteinase isolated from the filamentous fungus *Trichoderma reesei* using centrifugal ultrafiltration. The experiment rapidly creates a protein concentration gradient resulting in supersaturation and crystallization. I have used centricon-10 devices (Amicon) that allow solute to pass through the 10,000 MW cutoff membrane while retaining the protein, thus increasing its concentration. The purified enzyme was added to a pre-washed centricon-10 device at 0.5 mg ml⁻¹ in 0.1 M sodium acetate at pH 5.6 in a volume of 2 ml. The sample was reduced by centrifugation at 3,000*g* (Sorval RC-5B, SS-34 rotor) for 25 min at 4 °C to 0.5 ml and the spin-through buffer solution was removed. Another 2 ml of dilute enzyme solution was layered carefully onto the retained concentrated protein solution and centrifuged again. This process was repeated until 10 ml of the proteinase solution had been processed. The final protein concentration was 20 mg ml⁻¹

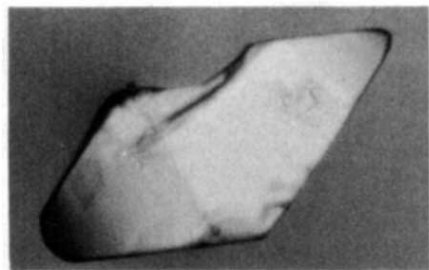


Fig. 1 A crystal of the aspartic proteinase from *T. reesei* grown in 2.5 hours by centrifugal ultrafiltration.

and the experiment took 2.5 hours.

During the removal of the concentrate, I observed large crystals; these were carefully transferred into a microfuge tube and stored at 4 °C. Figure 1 shows one of the crystals produced with approximate dimensions of 0.6 × 0.4 × 0.2 mm recorded on a Zeiss photomicroscope. I selected one of these for analysis

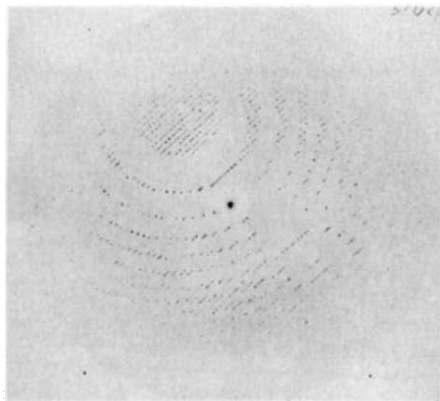


Fig. 2 A 1.5°-oscillation photograph of a single crystal of *T. reesei* aspartic proteinase grown by centrifugal ultrafiltration, indicating good diffraction to high resolution (2.3 Å).

at the SERC synchrotron X-ray source at Daresbury, UK; the preliminary result is shown in Fig. 2.

The rapid production of high-quality crystals using a low speed centrifuge and commercially available concentrators offers a new approach to the problem of crystallization. The forced increase in concentration of the enzyme under low centrifugal force while maintaining the starting buffer conditions allows this parameter to be varied singly and should produce a more controlled approach to supersaturation. The exploration of other protein systems by the use of small crystal seeds may allow a rapid improvement in crystal size. Additionally, information from screening trials indicating insolubility or the formation of microcrystals may provide suitable conditions for crystallization using this method.

Although the success in producing usable crystals for X-ray analysis in less than 2.5 hours indicates the potential of this approach, further investigations are needed to explore its general applicability.

JIM E. PITTS

Department of Crystallography,
Birkbeck College, University of London,
Malet Street, London WC1E 7HX, UK

1. Dover, S. D. *Nature* **354**, 388 (1990).
2. McPherson, A. *Preparation and Analysis of Protein Crystals* (Wiley, New York, 1982).
3. DeLucas, L. J. et al. *Science* **246**, 651–654 (1989).
4. Visuri, K. et al. *BioTechnology* **8**, 547–549 (1990).

SIR — We reported recently¹ that coal could be used as a source material to produce a crude fullerene mixture consisting of 90% C₆₀, 10% C₇₀ and trace amounts of other fullerenes. The separation of such a mixture into its components is an arduous task, usually carried out by chromatography on silica² or alumina³. We have found that fullerene mixtures from coal can be purified by chromatography on the same coal from which it is derived.

We ground and size-separated a sample of an Australian semi-anthracite, a coal from the Yarrabee mine, Queensland, (H/C ratio 0.74, O/C 0.013) known to be a good source of fullerenes⁴. The fraction that passed through a 125-μm-mesh sieve but was retained by a 63-μm-mesh sieve was separated and extracted in a Soxhlet apparatus for three days with toluene to remove soluble material. A glass column (1.2 × 58 cm) was packed with the 63–125-μm fraction using hexane, and a crude fullerene mixture (15 mg) was dissolved in toluene (3 ml) and adsorbed on the head of the column. After elution with hexane (800 ml) at a flow rate of 2 ml min⁻¹, 8.7 mg of solid was obtained from the eluate. This was identified as pure C₆₀ by infrared spectroscopy⁵⁻⁷. Further elution with 10% toluene/90% hexane mixtures (80 ml) did not yield distinctive magenta solutions of C₆₀, and further elution was continued with toluene. Initial elution with toluene (200 ml fraction) yielded C₆₀/C₇₀ mixtures enriched in C₇₀ relative to the crude fullerenes. Further elution with toluene (200 ml) produced fractions enriched in C₆₀, suggesting at least two different binding sites for C₆₀. Recovery of fullerene was quantitative within experimental error. The spent coal can be used for further purification of C₆₀/C₇₀ mixtures.

The ability of coals to separate fullerenes may derive from the disordered sheets of aromatic rings that they contain. Either the pores formed by these sheets may be the right size for separating fullerenes or there could be interactions between the sp² hybridized structures in coal and the fullerenes.

1. Pang, L. S. K., Vassallo, A. M. & Wilson, M. A. *Nature* **352**, 480 (1991).
2. Ajie, H. et al. *J. phys. Chem.* **94**, 8630–8633 (1990).
3. Taylor, R., Hare, J. P., Abdul-Sada, A. K. & Kroto, H. W. *J. chem. Soc. Chem. Commun.* 1423–1425 (1990).
4. Greenwood, P. F., Strachan, M. G., Willett, G. D. & Wilson, M. A. *Org. Mass Spectrometry* **25**, 353–362 (1990).
5. Hare, J. P. et al. *J. chem. Soc. Chem. Commun.* 412–443 (1991).
6. Vassallo, A. M., Pang, L. S. K., Cole-Clarke, P. A. & Wilson, M. A. *J. Am. chem. Soc.* **113**, 7820–7821 (1991).
7. Bethune, D. S. et al. *Chem. Phys. Lett.* **179**, 181–186 (1991).

We suggest that there is considerable potential for chemically modifying or choosing different coals for separation of different fullerenes.

M. A. WILSON
L. S. K. PANG
A. M. VASSALLO

CSIRO Division of Coal and Energy
Technology,
PO Box 136, North Ryde,
NSW 2113 Australia

Lacking a solution?

SIR — Magrath¹ points out that David Lack's well-known explanation of the clutch size of birds was anticipated by Eric B. Dunlop. Clutch size is, however, only a special case of the general problem of animal fecundity. This has a long and tortuous history.

R. A. Fisher, apparently unknown to Lack, discussed the general problem in his classic book² in 1930, hardly an obscure work. Fisher's characteristically incisive comments on fecundity may have been overlooked because they occur in the chapters dealing with human evolution and eugenics, from which many readers avert their eyes.

Fisher himself attributed the solution of the problem to his friend and patron Major Leonard Darwin, a younger son of Charles Darwin. Leonard Darwin did indeed cover the matter clearly³. He thought his treatment of the subject was novel, but he was wrong. Ironically, his father had discussed the problem in the first edition of his *Descent of Man*⁴ and given essentially the same solution. But in the second edition the relevant chapter was heavily revised and the passage on fecundity omitted. As later biologists seldom consulted the scarce first edition, Darwin's contribution seems to have been forgotten, even by his own son.

As a final complication, Charles Darwin gave the credit for the crucial insight to Herbert Spencer, who had discussed fecundity at length⁵. Spencer's treatment of the subject, like most of his writings, is not ideally clear, but the key point of 'Lack's solution' does seem to be there. Perhaps then it should be 'Spencer's solution' — unless an even earlier example comes to light.

DAVID BURBRIDGE

9 Penrith Road,
Thornton Heath,
Surrey CR7 8PN, UK

1. Magrath, R. D. *Nature* **353**, 611 (1991).
2. Fisher, R. A. *The Genetical Theory of Natural Selection*, 185–186 (Oxford University Press, 1930).
3. Darwin, L. *Eugenics Review* **14**, 266–269 (1922–3).
4. Darwin, C. R. *The Descent of Man, and Selection in Relation to Sex* Vol. 1 318–320 (Murray, London, 1871).
5. Spencer, H. *The Principles of Biology* Vol. 2 472–473 (Williams & Norgate, London, 1867).

Mirror-script and left-handedness

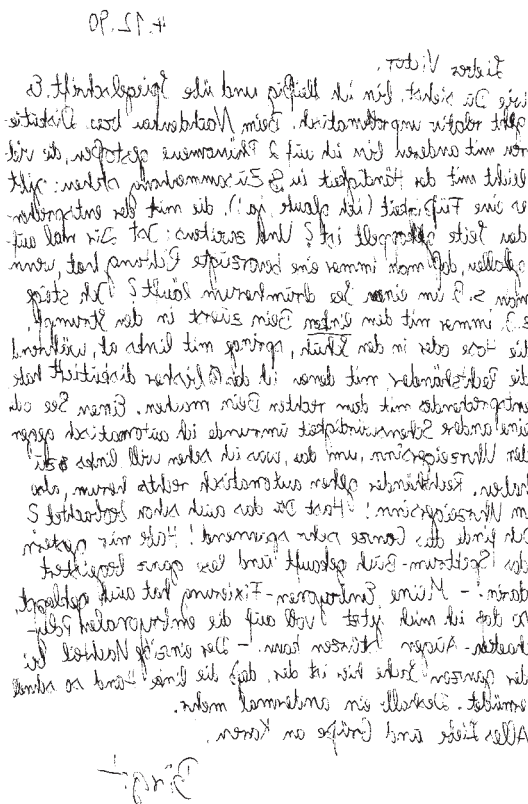
SIR — Leonardo da Vinci's famous sketches are accompanied by illegible notes which, viewed in a mirror, turn out to be neatly written Roman script. Reading and writing mirror script is very difficult for most people, so why did Leonardo go to the trouble? As a code, it is much too easy to crack, and if this

handwriting is reasonably neat. One of them (J. G.) wrote her thesis for a master's degree in mirror script on transparent paper and made corrections on the flip side with her right hand. She claims that the flow of thought finds better expression in words when writing mirror script with her left hand. She took recourse to this method following several frustrating attempts to write in normal script with her right hand. Both style and scientific quality improved significantly in the final, mirror-script version. I suggest that Leonardo had the same problem.

In naturally left-handed persons, language ability is located in the right hemisphere of the brain. Enforced right-handed writing means that sentences produced in the right hemisphere have to be transferred to the left hemisphere guiding the writing right hand. Apparently, a mirror version of the script is unconsciously anticipated in the right hemisphere before transfer to its left counterpart but can be directly transferred to the left hand. Not surprisingly, reading abilities are not affected, although Leonardo must have subsequently trained himself to read his own mirror-script. Transfer between hemispheres will be cumbersome and impede the free flow of thought. Indeed, suppressing left-handedness is known to lead to psychological problems such as stammering. Another person who fluently wrote mirror script was C. L. Dodgson, better known as Lewis Carroll, the author of *Alice in Wonderland* and, more significantly, *Through the Looking Glass* (C. Newman, *Natn. Geogr.* **179**, 100; 1991). Dodgson's peculiar ability to write nonsensical masterpieces, coupled with his stammering and reclusiveness, could have been due to suppression of natural left-handedness.

There is another, more significant dimension. Both the Etruscan and Archaic Roman scripts — the precursors of this script — were mirror versions, written from right to left. Could these early scripts originally have been written with the left hand, with the subsequent switch to the modern version occurring after writing with the right hand became customary?

Considering that most of the human population has always been right-handed, it is difficult to imagine that



A letter written to me 4 days after its author discovered her ability to write mirror script with her left hand.

ability is merely a further sign of the master's genius, why did he consistently use mirror script?

I suggest a simple explanation, based on the known fact that Leonardo was naturally left-handed. In accordance with the customs of the times, he may have been forced to learn to write with his right hand in his childhood and youth. Later in life he must have drawn with his left hand; there is good reason to believe that the script that flowed naturally from that hand would have been mirror script.

As late as the 1960s, German children were commonly forced to write with their right hands. I know four people who are clearly left-handed but who write with their right hands for that reason. All four experience difficulty in writing normal script with their left hands, but write mirror script fluently with the same hand (see figure). All were surprised upon discovering this latent ability. Reading their own script takes much longer than writing it although, when viewed in a mirror, the

writing was originally done with the left hand. The alternative explanation, that naturally left-handed persons invented phonetic scripts, is even more far-fetched. One possible avenue of exploration could be that phonetic scripts were conceived in the hemisphere (known to be associated with musical ability) opposite to Broca's area, finding their outlet via the left hand. The transition from left-handed to right-handed writing occurred gradually in ancient Greece, exemplified by several texts (such as on Apollo's temple at Delphi) written in zig-zag form with normal and mirror script alternating. As many Greek (and Roman) capital letters are bilaterally symmetrical, it is not difficult to read mirror script. By contrast, establishing mirror versions of asymmetrical scripts such as Semitic would have been tantamount to creating a new script, which is perhaps why the original version was retained after the more dexterous hand took over writing. Indeed, the practice of suppressing left-handed writing might well have had its origins in an effort to establish a single writing style rather than in a mystical aversion to the left hand, as is widely believed.

VICTOR SMETACEK

Alfred Wegener Institute for Polar
and Marine Research,
2850 Bremerhaven, Germany

Virus release evaluation

SIR — There have been several recombinant approaches to increase the speed by which insect-specific viruses kill pest species; recent data indicate that insect viruses can be engineered to increase dramatically the speed of kill¹⁻⁴. This advance makes these viruses far more attractive for use in agriculture. Williamson's Scientific Correspondence⁵ addressing the News and Views article by Hochberg and Waage⁶ is valuable in focusing attention on the safety of these genetically engineered organisms. Although these prototype viruses offer numerous potential benefits⁶, possible risks need to be addressed, certainly by discussions between scientists and the public, and by design of rigorous experiments to evaluate safety.

These viruses are being designed as insecticides, not as biological control agents that will become permanently established. As augmentative biological control relies increasingly on inundative release of commercially available agents, the strategies being developed tend to be like those for insecticides. In such releases it is important to adopt a more holistic approach cognizant of environmental consequences, not only for natural and engineered pathogens,

but also for predaceous and parasitic arthropods^{7,8}. By referring both to the wild-type and engineered viruses as viral insecticides, we remind ourselves not to repeat the mistakes that were made with synthetic chemical insecticides.

This terminology also has a positive connotation regarding Williamson's legitimate worry that nontarget lepidopterous larvae could be permanently endangered by these engineered viruses. Because these materials are being designed for an insecticide-type application and not as 'biological control' agents that will replicate in the environment, this concern is in part overcome. The fact that these viruses kill insects much more quickly than wild-type viruses leads to far fewer infective virus particles being produced, because the larvae are killed when they are very small and at an early stage in viral replication. Thus, by their mechanism of these actions, recombinant viruses will be less competitive than wild-type organisms. The risk to nontarget species seems more analogous to classical synthetic insecticides in that it arises primarily from the initial application. Recombinant viruses appear to vanish rapidly from the ecosystem, and the risk to nontarget species following direct application of large doses is being systematically addressed by the Oxford NERC group referred to by Williamson. Data so far indicate an intrinsically greater specificity of the recombinant viruses for target insects than either synthetic chemical insecticides or *Bacillus thuringiensis*. It is critical that such studies on stability and specificity of the viruses continue.

Certainly, it is important to disable the viruses. But the approach of using polyhedron-negative constructs as suggested by Williamson results in viruses so severely disabled that they are of little value in agriculture. Fortunately, there are other means to limit replication of engineered viruses in the field. Further research into the molecular mechanism of host specificity will enhance our ability to target more pest species, with greater safety to nontarget species.

BRUCE HAMMOCK

Departments of Entomology and
Environmental Toxicology,
University of California,
Davis, California 95616,
USA

1. Tomalski, M. D. & Miller, L. K. *Nature* **352**, 82-85 (1991).
2. Stewart, L. M. D. *et al.* *Nature* **352**, 85-88 (1991).
3. Maeda, S. *et al.* *Virology* **184**, 777-780 (1991).
4. McCutchen, B. F. *et al.* *BioTechnology* **9**, 848-952 (1991).
5. Williamson, M. *Nature* **353**, 394 (1991).
6. Hochberg, M. E. & Waage, J. K. *Nature* **352**, 16-17 (1991).
7. Ehler, L. E. in *Assessing Ecological Risks of Biotechnology* (ed. Ginzburg, L. R.) 21-39 (Butterworth-Heinemann, New York, 1991).
8. Ehler, L. E. *Envir. Ent.* **21**, 1-15 (1992).

Mongols and soap

SIR — Cowart¹ has recently explained how to test the clinical disorders of smell. The earliest record of a disturbing soap-eating experience appears in the Thirteenth century, when the Mongol envoys came to the court of French king Philip *Le Bel* with the goal of discussing the joint Crusade in the Middle East². Upon arrival, the Mongol ambassadors were offered a bar of soap and a towel to wash off the dust from their long journey. The Mongols, assuming that the presentation of a piece of dried cheese was also a customary homage in the land of the Franks, had a good bite of it. It is unlikely that the Mongols were anosmic. It is also questionable that Philip IV was a notorious olfactory psychophysicist and was conducting chemosensory measurement experiments.

We do not know of the fate of the washroom attendant at the hands of sixteen deceived Barbarians whose tribe had just conquered half of the world. There is little doubt that the Mongols preferred to conclude the incident by demonstrating their skill in martial arts rather than in diplomacy. The influence of this episode on the gloomy destiny of the Crusades is not yet measured, but since then *savon* became the only known French word adopted by the Mongol vocabulary.

A. S. BOURINBAIAR

118 E 110th Street,
New York, New York 10003, USA

1. Cowart, B. J. *Nature* **354**, 114 (1991).
2. Morgan, D. *The Mongols* (Blackwell, London, 1986).

Diabetes and heat shock protein

SIR — Cornall *et al.*¹ identify a new susceptibility gene, *IDD5*, for insulin-dependent diabetes in the non-obese mouse. The identity of the gene product in this region, however, remains unknown, although linkage has been identified to the interleukin-1 receptor and to the *Lsh/Itg/Bcg* gene. This gene is involved in the immune response to several organisms such as mycobacteria in which the major antigen involved is heat-shock protein 65. The human equivalent of *IDD5* seems to be located within the 2Q32 region¹.

We have previously identified human heat-shock protein 65 as being a β -cell antigen in insulin-dependent diabetes in man and have identified the presence of binding antibodies to this antigen in the serum of a large proportion of newly diagnosed insulin dependent diabetic patients². We have used a human heat-shock protein 65 gene probe λ C5 to map the location of this gene by fluorescence *in situ* hybridization with a biotin-

labelled probe³. We obtained hybridization signals at multiple sites — 2q32–33, 5p14, 8p23, 12p13 and 12q13, suggesting several HSP65 loci, some of which may be pseudogenes. Thus IDD5 may represent the mouse equivalent of a heat-shock protein 65 gene. The expression of β cell heat shock protein 65 and its genetic control may be involved in the development of insulin-dependent diabetes mellitus.

DAVID BOWEN-JONES

Consultant Physician,
Arrowe Park Hospital,
Upton, Wirral L49 5PE, UK

JUDITH FANTES

MRC Human Genetics Unit,
Western General Hospital,
Edinburgh EH4 2XU, UK

RADHEY GUPTA

Department of Biochemistry,
MacMaster University,
Ontario L8N 3ZS,
Canada

1. Cornall, R. J. *et al. Nature* **353**, 262–264 (1991).
2. Jones, D. B. *et al. Lancet* **336**, 583–585 (1990).
3. Jindal, S. *et al. Molec. cell. Biol.* **9**, 2279–2283 (1989).

Amphibian decline?

SIR — Implicit in the recent discussion about global declines of amphibians is the suggestion that this class of vertebrates is at a greater risk than others

Kuwaiti soot over Japan

SIR — We carried out an aircraft observation on 27 April 1991 over Tsukuba, Japan. Individual aerosol particles collected at 7.5-km altitude were examined using an electron microscope with an energy-dispersive X-ray analyser. We have found an extremely high proportion (53%) of soot-containing particles in the submicrometre-size range (0.15–1- μ m radius). Such a high fraction of soot particles in the upper troposphere has not been reported previously.

In the normal upper troposphere over Japan, sulphuric acid and/or ammonium sulphate particles are dominant. In our latest samples, we found some dust-storm particles from China, but soot particles were not mixed with this dust. Thus, we consider that the soot particles we have collected do not come from East Asia and are due to long-range transport. Because vanadium was detected in some of the soot particles, we believe that the soot particles from the upper troposphere originate from the combustion of oil.

Soot particles have been emitted from several hundred oil-well fires in Kuwait since February 1991 with large emission rates of approximately 5 Tg yr⁻¹ (refs 1, 2). Aircraft observation over the Gulf in March 1991 showed that significant amounts of smoke were present only below 5-km altitude². However, in

cloudy cyclonic situations, trace amounts of the smoke may reach regions of 8–12-km altitude³ and the stratosphere⁴. Soot particles from Kuwait could be transported vertically by deep convective clouds and influence the features of aerosol particles in the upper troposphere, where aerosol concentrations are usually low. Thus, the origin of these soot particles over Japan is very likely to be oil-well fires in Kuwait.

From the backward- and forward-trajectory analyses of air particles obtained by the techniques⁵, the soot particles over Japan are expected to be supplied to the upper troposphere in the first half of April and those have circled the globe before arrival at the sampling point.

KIKUO OKADA

MIWAKO ILEGAMI

OSAMU UCHINO

YOSHINOBU NIKAIKIDOU

YUJI ZAIZEN

YUKITOMO TSUTSUMI

YUKIO MAKINO

Meteorological Research Institute,
1-1 Nagamine, Tsukuba,
Ibaraki 305, Japan

1. Small, R. D. *Nature* **350**, 11–12 (1991).
2. Johnson, D. W. *et al. Nature* **353**, 617–621 (1991).
3. Browning, K. A. *et al. Nature* **351**, 363–367 (1991).
4. Bakan, S. *et al. Nature* **351**, 367–371 (1991).
5. Nikaidou, Y. *et al. in Research Activities in Atmospheric and Oceanic Modelling* (ed. Boer G. J.) **263**, 6.12–6.13 (WMO/TD, 1988).

DISTRIBUTION AND STATUS OF FRESHWATER FISH, AMPHIBIANS AND REPTILES IN EUROPE

Class	Total no. species	Species found in <5% of area	Species found in >25% of area
Freshwater fish	130	35 (27%)	52 (40%)
Amphibians	44	18 (41%)	17 (39%)
Reptiles	82	39 (48%)	11 (13%)

	Species			
	On Appendix II of Bern Convention	In Red Data Book	On CITES appendices	Considered endangered ⁴
Amphibians	17	5	0	6
Reptiles	34	4	8	17

(see refs 1, 2).

But the evidence for this assertion is patchy, relating mainly to a selection of particular species in northern latitudes or upland regions of the world. Because the distribution and status of amphibians and reptiles in Europe are well documented (see refs 3–5), I have made a systematic comparison of the two

groups using this published information.

The position with respect to species diversity and official perceptions of risk (as determined by the Bern Convention, CITES appendices, the *Red Data Book* and ref. 4) is summarized in the table.

Although amphibians are the least diverse vertebrate group in Europe and the rest of the world, reptiles are in a worse position by several criteria: a higher proportion of reptile taxa have small overall ranges, and far fewer have large ones; eight European reptiles, but no amphibians, feature on CITES appendices; and a higher proportion of reptile species were recently considered endangered⁴. Only *Red Data Book* classification implies a relatively worse situa-

tion for the Amphibia. Further analysis (by multiple regression) shows that both amphibian and reptile declines (measured as per cent endangered and vulnerable in 22 European countries) are relatable to latitude and human population density in highly significant ways.

Both amphibians and reptiles have declined in Europe recently, but the average proportion of endangered and vulnerable reptile species in European countries (32%) is marginally higher than that of amphibians (27%), and neither group has suffered a single extinction within historical times. Both groups need to be conserved, but this will be best achieved by improved habitat protection and pollution control rather than by pursuit of class-specific extinction factors for which, in Europe, there is no substantive evidence.

T. J. C. BEEBEE

School of Biological Sciences,
University of Sussex,
Falmer, Brighton BN1 9QG, UK

1. Blaustein, A. R. & Wake, D. B. *Trends Ecol. Evol.* **5**, 203–204 (1990).
2. Wake, D. B. *Science* **253**, 860 (1991).
3. Arnold, E. N., Burton, J. A. & Ovenden, D. W. *A Field Guide to the Reptiles and Amphibians of Britain and Europe* (Collins, London, 1978).
4. Corbett, K. F. (ed.) *Conservation of European Reptiles and Amphibians* (Helm, London, 1989).
5. Groombridge, B. *World Checklist of Threatened Amphibians and Reptiles* (Nature Conservancy Council, 1988).

Scientific Correspondence

Scientific Correspondence is intended to provide a forum in which readers may raise points of a scientific character. They need not arise out of anything published in *Nature*. In any case, priority will be given to letters of fewer than 500 words and five references. □

Market research

Robert M. Rosenzweig

The Changing University: How Increased Demand for Scientists and Technology is Transforming Academic Institutions Internationally. Edited by Dorothy S. Zinberg. Kluwer: 1991. Pp. 182. £41, \$69.

THE bewildering pace of world events may yet turn today's wave of the future into tomorrow's beached driftwood. But for now it seems that market forces are widely accepted as the most effective means of regulating economic activity. And it is not hard to understand why. The prospect of reward proportionate to risk and earned through competition can unleash enormous energy. The dynamism of the world's free economies stands in sharp contrast to the stagnation of the failed economies of the Socialist nations.

Systems of higher education could hardly stand apart from so profound a force. In showing how and why they have not, this book addresses whether market forces are on balance likely to be helpful or harmful to universities and science. Even in the United States, where universities have never really been isolated from the rest of the world, the pressures and temptations of closer connections with commerce have introduced policy issues that are testing the fabric of these institutions. The value of the book is that it examines in an international context issues already familiar to US and, increasingly, British academics. It provides a useful guide to these pressing issues, a sobering reminder of their capacity to do damage, and an invitation to remember first principles.

The essays can be grouped under three main themes: changing patterns of funding and their consequences; science and engineering manpower; and the effects of university involvement with government and industry on the free flow of scientific information. In all three areas, the news is not good for those who hold to the classical view of the university both as an institution committed to the search for truth and to the dissemination of this knowledge without hindrance, and as a body opposed to all forms of parochialism. It is clear that in the United States and Europe at least, the perceived economic value of research and training in science and technology has produced a market for what universities have to offer, and even those institutions most faithful to their traditions are finding it hard to resist the market's claims on their resources and policies.

It is useful to view these matters from an industrial perspective. In a thoughtful essay that is largely sympathetic to the

university cause, John A. Armstrong of IBM raises the uncomfortable questions that university connections with business inevitably present. He clearly spells out the harsh reality that industry has no obligation to support universities beyond the support that industrial taxes provide. He believes that it is unrealistic to expect industry to make up for shortfalls in government funding, adding that it "would be a mistake to ignore the substantial ignorance and lack of sympathy, both in corporations and in government for the ethos of the open, internationally minded university community". More specifically, he argues that universities have greatly overstated the value of their research to economic competitiveness; that important conflicts over the ownership of intellectual property remain unresolved; and that universities allow and even promote profit-seeking conduct that is unacceptable for scientists in industrial or government laboratories.

Shirley Williams offers a British per-

spective, but as the main aim of recent British governments seems to have been to make British universities more like those in the United States, her words really address both systems. They are words of caution: "Those who want to harness the universities to commercial objectives may destroy the very qualities they admire in them — intellectual excellence, free inquiry, scientific imagination."

The evidence that such warnings need to be taken seriously is, not surprisingly, most readily seen in the United States, where universities are the most permeable by commercial enterprise. It would be an exaggeration to say that there is a crisis of values in US universities, but not one to say that the elements of such a crisis exist and will grow worse unless attended to. Universities now routinely accept the need to delay the publication of scientific work in order to protect patenting or other proprietary rights. Faculty consulting and entrepreneurial activities compete with other institutional obligations.

Sharp criticism of the willingness of universities to accept foreign, especially Japanese, research and other support conflicts with the growing need to make education and science more international. So, too, do concerns over the number of foreign students in science and engineering doctoral courses, the



HUNTERS in Yosemite, a decade before the region was accorded national park status in 1891. The mountaineer and popular nature writer John Muir welcomed the restrictions on "the barbarous slaughter of bears, and especially of deer", although the interests of the visitors came first, with aggressive bears being sometimes 'disposed of'. Recreational hunting remained a widely accepted activity, endorsed by conservationists and influential statesmen such as President Theodore Roosevelt, who hoped that wilderness hunting would restore to urban Americans the "hardihood, self-reliance, and resolution" of the pioneers. Following a hunting trip in Yosemite in 1903, he was scolded by Muir who asked, "Mr. Roosevelt, when are you going to get beyond the boyishness of killing things?" In *Wild Animals and American Environmental Ethics*, Lisa Mighetto draws on the writings of these and other American conservationists to trace the changing attitudes to wildlife in the United States since the mid 1800s. Published by The University of Arizona Press, price \$35 (hbk), \$17.95 (pbk).

increasing dependence of the science and engineering workforce on those students when they graduate, and the possibility that many of them will return to their more prosperous homelands to become the spearhead of further competition to US industry. The modern equivalent, some would say, of capitalists selling rope to their hangmen.

The world will not change because a few academics are worried that the way it is moving may damage their institutions. Still, there is nothing inevitable about the direction of events, and academics and their friends are not wholly excluded from debates about policy in most countries. Here are some suggestions for constructive action.

First, at every opportunity, remind economists and their allies in government that markets do only one thing very well, namely, allocate resources by setting prices. Markets are not automatically applicable to all social problems, and they may be fatally destructive where values are at issue and quality matters more than price.

Second, welcome the opportunity to be useful. It is not a crime to concede that society has a right to a reckoning of the value of its academic institutions and to expect that some of their intellectual power should address today's problems. As Derek Bok points out in his essay,

there is little evidence to support a direct connection between research and the economic prosperity of the country in which it takes place. In the United States, the connection has already been oversold. The argument needs to be refined and moderated to prevent universities from becoming swamped by disappointments.

Finally, academics need first to understand themselves, and then to persuade others that their greatest social contribution, by far, lies in the preservation of the values that make their institutions worth supporting in the first place. Those values are important underpinnings of democratic societies, as well as being essential for high-quality teaching and research. □

Robert M. Rosenzweig is president of the Association of American Universities, 1 DuPont Circle NW, Washington DC 20036, USA.

■ Robert M. Rosenzweig is one of 14 contributors to *Research and Higher Education: The United Kingdom and the United States*, edited by T. G. Whiston and R. L. Geiger. The book considers the 'national systems'; research environments, teaching and research interactions, and graduate education and research training; and prospects for academic research in relation to governments, industry and society. Published by Open University Press, £35.

Mixed reactions

Mark Varney

ECos Estuarine Contaminant Simulator. *Plymsolve*, 32 Looe Street, Plymouth PL4 0EA, UK: 1991. Licence £560 or £280 (educational and research). Other rates and details on application.

ESTUARIES, with their high productivity and frequent proximity to cities, are the purview of marine scientists, teachers, environmentalists and all those involved in aquaculture, water monitoring and pollution control. Many of these groups now use computer software tools to provide a better understanding of data, to assess and predict impacts and risks and to assist in writing reports. These tools are valuable provided that their limitations are acknowledged; they will certainly never obviate the need for field studies. ECoS, a menu-driven estuarine contaminant simulator, is the latest addition to these aids and is intended for specialists and nonspecialists alike.

A main concern of estuarine science is to calculate the transport of contaminants from rivers through estuaries and into oceans. Direct measurements of the hydrodynamics, water flow and concentrations of contaminant involved in the transport are difficult and expensive to

make — hence the need to model and predict fluxes and distributions.

Obviously, modelling any estuary is immensely difficult — tides, the weather, intertidal and mudflat zones, waves, hydraulics, dredged channels and shipping all influence the movement of water. ECoS contains a segmented physical mixing model that can account for the transport and dispersal of contaminants by river flow and tides, water solubility, degradation, disequilibria, air-sea exchange, partitioning onto suspended particles and sediment interactions. The attraction of the simulator is that it can be built up without any knowledge of the underlying physics or hydrodynamics.

The software is written as a flexible 'shell' that can be tailored to both real and hypothetical estuaries. The shape of the estuary can be represented algebraically or as a series of cross-sectional 'segment' areas constructed from 'boxes'. One can define exactly how and where in the estuary the contaminants are introduced, react and are lost, and the program can then be run to establish the distributions of dissolved and absorbed contaminants at a later stage, as well as calculating their residence times in the estuary. Parameters can be freely altered so that their effects on the distributions can be observed. The program can be run and re-run with slight modifications to the 'forcing' functions,

allowing the importance of other chemical processes to be examined interactively. Simulations can be saved and replayed, with the addition of new data if necessary. In this way, large libraries of data and results can be collected.

What are the program's limitations? Measuring fluxes in estuaries is difficult because of the variability caused by external factors. ECoS therefore needs to be calibrated to validate the physical-exchange equations, requiring exacting surveys, usually of the salinity gradient as the best available tracer representing the mixing processes.

ECoS, like most models, uses the averages of measurements throughout tidal cycles. So 'double high waters' and 'young flood stands', which may have significant and asymmetric effects on sediment transport, cannot be modelled. ECoS also assumes that each 'segment' in the model is well mixed. The model cannot represent side-to-side exchange processes such as wind-driven effects on the mean circulation, which in any case are difficult to measure accurately, as are freshwater flows independently of tidal influences. These inaccuracies can lead to large error in flux estimates. And tidal range, river discharge and wind can affect fluxes for up to hundreds of tidal cycles, influences that must be considered when assessing the general applicability of a particular result. Further, with ECoS one cannot model layering such as that of salt wedges or the freshwater-saltwater interface, except by the difficult procedure of increasing the number of 'boxes' where the concentration gradient is highest. Nor does the program treat the problem of 'embayment' — instead, boxes have to be added to the side of the main estuarine box set.

The program is easy to install and use, although it is notably slow and not readily compatible with some other programs, particularly those that are memory-intensive or use a mathematics coprocessor. A code from the hard disk is used as a password to protect the program from illegal copying (I would have preferred the use of the 'dongle').

ECoS is not an ideal representation of an estuary. Two-dimensional models usually provide better descriptions but are more difficult to write, require greater computing power, and are often specific to a particular location and cannot be readily modified. ECoS is useful because it can help determine if it is worth developing better models. It will prove to be valuable for quick exploratory analysis and for indicating where it would be fruitful to collect more data. □

Mark Varney is in the Department of Oceanography, University of Southampton, Highfield, Southampton SO9 5NH, UK.

Worlds apart

Ann M. Clarke and A. D. B. Clarke

Autism and Asperger Syndrome. Edited by Uta Frith. Cambridge University Press: 1991. Pp. 247. £30 (hbk), £10.95 (pbk).

IN 1943, Leo Kanner, an *émigré* Viennese psychiatrist and then head of the Johns Hopkins Clinic in Baltimore, published an article entitled "Autistic disturbances of affective contact" which was to become famous worldwide. In it he introduced the label "early infantile autism" for a type of severe, pervasive disorder previously unrecognized as a clinical entity, although earlier cases are now known to have existed. Other cases rapidly came to light and today it is thought that between 4 and 10 children per 10,000 are affected. Dustin Hoffman recently brought autism to the attention of cinema audiences in *Rain Man*.

Just one year after Kanner's contribution, there was a remarkable coincidence when another Viennese psychiatrist, Hans Asperger, separated by the Second World War from US literature, published a paper entitled "'Autistic psychopathy' in childhood". He reported, among other things, oddities in language, lack of humour and pedantry, serious attention deficits, clumsiness, avoidance of eye contact, strange gestures and posture, and specific learning difficulties, even in brighter children.

Kanner believed that the problem involved odd relations with people only, but Asperger had a wider concept — that the deficit related not only to people, but also to objects. As Uta Frith indicates, Kanner's syndrome is associated with children who do not talk or who 'parrot' speech and use strange phrases, who line up toys in long rows, are oblivious to other people and remember meaningless facts. On the other hand, typical of Asperger's syndrome are children or adults who are socially inept but often socially interested, articulate yet strangely ineloquent, markedly gauche and impractical, and specialists in unusual and often narrow fields. Is Asperger's syndrome a form of autism? This question forms the focus of this volume.

Frith provides an annotated translation (the first into English) of Asperger's original book and brings together a small group of seminal writers in the field of autistic disorders. She introduces the paper with a chapter on Asperger and his syndrome, arguing that affected individuals suffer from a particular, and not especially rare, form of autism. A Swedish survey of a large sample of children indicated a prevalence for the syndrome

of between 10 and 26 per 10,000, at least double the number for classic autism. Firth suggests that autism is a developmental disorder and thus its behavioural manifestations will vary with age and ability. Its main features, however, can take different forms. Lorna Wing proposed a triad of core symptoms — impairment in socialization, communication and imagination. Affected persons do indeed have these impairments, and

syndromes, argues for a continuum of impairments within autistic disorders. An autistic child of normal intelligence may miss being diagnosed, but may require psychiatric help in adulthood. To us this suggests that there is a much larger group of 'loners' with a 'touch' of autism who, with family support, evolve strategies to cope, such as that of obsessive and narrow interests, and who remain as eccentrics in the community, without needing special help.

Christopher Gillberg presents six detailed family studies, illustrated with genealogical trees, each containing a remarkable distribution of members with Asperger traits, autism or full Asperger's syndrome. These studies illuminate various clinical and neurobiological issues, and suggest that clumsiness may possibly distinguish Asperger's syndrome from autism. Digby Tantam adds to the overall picture with a summary of two large descriptive studies of adults with Asperger's syndrome.

The inclusion of two chapters by people who are closely involved with individuals having Asperger's syndrome, or who present personal accounts, brings this book even more to life than do earlier case histories. Margaret Dewey reflects her knowledge as a parent of an affected child, but she is also a counsellor and provides general support to people with this handicap. Francesca Happé provides excerpts from three autobiographical pieces

by Asperger's sufferers, one, Temple Grandin a well-known expert on animal husbandry, who has published over 200 articles in the field.

This book is highly recommended to a variety of readers, including medical and behavioural scientists, parents, friends and even sufferers from Asperger's syndrome. It is clear, succinct and so far unique in its presentation of important findings relating to this impairment. There are many suggestions as to the next steps in empirical investigation, and a wealth of humane wisdom pervades the whole. □

Ann M. Clarke (Department of Education) and A. D. B. Clarke (Department of Psychology) are emeritus professors at the University of Hull, Hull HU6 7RX, UK.

■ *Molecular Genetic Approaches to Neuropsychiatric Diseases*, edited by J. Brosius and R. T. Freneau, covers diseases such as Alzheimer's and schizophrenia. Academic Press, price £69.95.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Advertising autism — Dustin Hoffman in *Rain Man*.

like other autistic individuals, they tend to show a typical pattern of performance in IQ tests. Unlike autistics, however, they usually score in the average range of intelligence, and some are very bright indeed.

Biological causes of autism seem to be manifold. Frith is associated with a theory of autism that leads to specific, experimentally verifiable predictions. Her approach suggests that affected individuals have inordinate difficulty in understanding other people's motives, beliefs and attitudes, making the world incomprehensible. They lack a "theory of mind", an ability to understand the minds or intentions of others, which is the basis of normal social intercourse. It seems that individuals with Asperger's syndrome are less severely impaired in the core symptoms, and their "theory of mind" is relatively intact.

Lorna Wing writes on both the similarities and differences between the two

B. F. I. Stills, Posters and Designs

Insights into setting sites

Peter Aldhous

The Fall-Safe Society: Community Defiance and the End of American Technological Optimism. By Charles Piller. Basic Books: 1991. Pp. 240. \$20.

NOT in my backyard! A new breed of environmental campaigner, the NIMBY activist, threatens the world's foremost technological society with paralysis. Across the United States, plans to build new waste dumps, factories and even basic research laboratories are being blocked by local residents. Agricultural biotechnology moves at a snail's pace, restrained by opposition to field tests of genetically engineered products; no new hazardous-waste dumps have been sited in a decade. To US technocrats, the picture painted by Piller must be depressingly familiar. In his book, billed as "the first balanced assessment" of the NIMBY phenomenon, Piller attempts to define the NIMBY activist, and to present US technological managers with a solution to a seemingly intractable problem. He can claim a degree of success in both of these tasks.

The book's core is a journalistic exposition of three case studies in NIMBYism: the opposition to the Department of Energy's Rocky Flats nuclear weapons production plant near Denver, Colorado; the controversy surrounding the first field trials of a genetically engineered bacterium; and the battle of the University of California at San Francisco (UCSF) with the residents of the affluent Laurel Heights suburb over plans to expand the university's biomedical research facilities. Piller provides lively accounts, interspersing the chronology of events with revealing profiles of the key players.

Definitions

Piller quickly makes the distinction between true NIMBY groups, usually *ad hoc* coalitions that dissolve once the battle against a perceived threat is over, and their occasional allies, the conventional environmentalists. He observes well the uneasy relationship between the local opponents to Rocky Flats, who care little about nuclear weapons so long as their gardens are free from plutonium, and the outside environmentalists and peace campaigners who share their protest against the plant.

Piller aims to show that the NIMBY response is remarkably consistent, and to dispel the technological establishment's criticism that NIMBY activists are modern luddites, ignorant of the

technology they oppose and motivated entirely by selfishness. For this purpose, the case studies are well chosen. (But Piller's guide to the NIMBY movement would be more satisfying had he given more coverage to the 1978 Love Canal incident, which he identifies as the birth of NIMBYism as a political force. There, militant local opposition to chemical pollution led to the declaration by the Carter administration of the town in New York State as a federal disaster area.)

The Rocky Flats plant, at which many locals were employed, was viewed favourably by most of its neighbours for nearly four decades. But the scrutiny that followed a 1987 plan to open a new waste incinerator at the site revealed contamination of both the site and its groundwater with plutonium, and flagrant breaches of environmental regulations. Rocky Flats was closed by the Department of Energy in November 1989 and has yet to reopen.

In comparison, the dangers posed by UCSF's proposed expansion of its biomedical research facilities were trivial. But residents of Laurel Heights were nonetheless able to use environmental legislation to halt the university's plans to install laboratories in a vacated insurance-company building.

Completing the picture is 'ice-minus', a recombinant *Pseudomonas* bacterium lacking the genes for proteins that act as nuclei for the formation of ice crystals and therefore able to protect crop plants from frost damage. Although thought to be environmentally benign by biologists (similar *Pseudomonas* deletion mutants exist in nature), ice-minus, being a recombinant strain, nevertheless alarmed farmers in the Californian counties chosen for field tests. Geneticists found their plans blocked for more than three years; field tests eventually went ahead, with no apparent environmental problems.

As Piller shows, ignorance is not the cornerstone of NIMBYism. When NIMBY organizers delve into the technological details, their protests become more intense. Self-reliance, distrust of the technological establishment, and a desire to control decision-making emerge as NIMBYism's distinguishing marks. But that is as far as the parallels between the three outbreaks of NIMBYism go. The disruptive tactics and occasional factual distortions employed by the Rocky Flats opponents are defensible, their actions seemingly motivated by a legitimate instinct for self-preservation rather than by selfishness. But George Carr, and the other members of the Laurel Heights Improvement Association, deserve less sympathy.

Piller is so keen to prove the consistency of the NIMBY response, that he

neglects an obvious interpretation of the actions of Laurel Heights' residents. Although the legal challenges to UCSF's plans were based on the possible danger of chemical and biological emissions, Carr originally complained about the disruption that an influx of delivery vehicles would cause in a cosy upper-middle-class neighbourhood, and about the possibility that meetings of the University of California Board of Regents would attract student protesters. One suspects that the environmental safety objections were always a secondary concern, but allowed Carr to play on public fears, and gave him the legal leverage needed to upset UCSF.

Solutions

Piller's second task, finding a solution to the problem, is more formidable. NIMBY attitudes stem from a feeling of powerlessness, according to Piller. If the energy of NIMBY groups is to be harnessed for society's use, he argues, local people must become involved in technological decision-making. Environmental safety standards need to err more on the side of caution; full disclosure about development plans and their potential hazards is essential, and must reach the grass roots, not just a handful of existing local officials; trained community representatives should be responsible for enforcing agreed environmental performance standards, once a project is agreed. All this needs money, says Piller, including government grants to enable community organizations to muster the scientific expertise for their contribution to the debate.

Piller's recipe has much to commend it, but there are obstacles (and not just those of cost): the end of the Cold War notwithstanding, secrecy will remain a preoccupation of those responsible for the US nuclear-weapons complex. Piller dismisses the potential of financial compensation to placate communities asked to host toxic-waste dumps and the like. But one intriguing idea is not discussed. This involves an auction where the figure offered in compensation rises until one of the communities short-listed for the siting of a potentially hazardous facility accepts the offer (see *Nature* 347, 611; 1990). Superficially at least, this 'reverse Dutch auction' meets Piller's requirement of giving local people greater control over their own destiny.

Piller may not have all the answers, but his book is a thought-provoking starting point for an important and necessary debate. As he notes, the days of 'decide-announce-defend' as the standard mechanisms for siting locally unwanted facilities are over. □

Peter Aldhous is a Nature staff reporter.

Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life

Christopher Chyba & Carl Sagan

Sources of organic molecules on the early Earth divide into three categories: delivery by extraterrestrial objects; organic synthesis driven by impact shocks; and organic synthesis by other energy sources (such as ultraviolet light or electrical discharges). Estimates of these sources for plausible end-member oxidation states of the early terrestrial atmosphere suggest that the heavy bombardment before 3.5 Gyr ago either produced or delivered quantities of organics comparable to those produced by other energy sources.

MICROSCOPIC fossils¹ and fossil stromatolites² indicate life originated on Earth more than 3.5 Gyr ago. Evidence for biologically mediated carbon isotope fractionation³ suggests that life may already have existed by 3.8 Gyr ago. The terrestrial origins of life must therefore have coincided with the final stages of the heavy bombardment of the inner Solar System, during which those planetesimals 'left over' from planetary formation were largely swept up or scattered. This bombardment^{4,5}, known from radioactive dating of cratered surfaces on the Moon, and by comparisons of the lunar, martian and mercurian cratering records, declined in intensity through orders of magnitude, reaching its present comparatively low level by ~3.5 Gyr ago.

The heavy bombardment seems to have had important consequences for the origins of life, some deleterious and some favourable. Sufficiently large and fast impacts can erode planetary atmospheres. This effect, although possibly critical for Mars⁶, was probably of little importance for the Earth⁷. The largest impactors could have led to an 'impact frustration' of life's origins⁸⁻¹⁰, through the creation of a globe-encircling rock vapour atmosphere and evaporation of the euphotic zone, or even the entire terrestrial ocean¹⁰.

During the heavy bombardment, volatile-rich impactors would have been delivering essential 'biogenic' elements to the terrestrial surface^{7,11-13}. Moreover, comets, carbonaceous asteroids, and interplanetary dust particles (IDPs) are rich in organic molecules¹⁴⁻¹⁶, so may have contributed directly to terrestrial prebiotic inventories. Impacts would also have shock-synthesized organics in the atmosphere^{17,18}. Here we focus on these last two effects, comparing them quantitatively with the principal non-heavy-bombardment sources of prebiotic organics.

Delivery of intact exogenous organic matter

Exogenous sources deliver organic molecules more or less intact to Earth today. These include¹⁴ those interplanetary dust particles small enough to be gently decelerated by the atmosphere, and meteorites large enough to avoid complete ablation, but small enough to be substantially decelerated during their fall. Some impactors catastrophically fragment during atmospheric passage, as seems to have happened^{19,20} to a comet or comet fragment over Tunguska, Siberia, in 1908. Fragmentation of a CI carbonaceous chondrite took place over Revelstoke, Canada, in 1965; photomicrographs of recovered millimetre-sized fragments reveal unheated interiors²¹, within which organics should have survived. (An investigation, probably requiring the examination of carbon isotope ratios, of whether exogenous organics are in fact present in the Revelstoke fragments needs to be done.) Finally, the discovery of apparently extraterrestrial amino acids in Cretaceous/Tertiary (K/T) boundary sediments at

Stevns Klint, Denmark, has been taken to suggest that a large fraction of cometary organics might in fact survive giant impacts²², although both xenon measurements at the K/T boundary¹⁴ and hydrodynamic simulations of organic pyrolysis in impacts¹⁵ argue strongly otherwise. Zahnle and Grinspoon²³ have invoked the accretion of cometary dust as an explanation; other explanations may also be possible²⁴.

Each of these exogenous sources of organics should have been present on early Earth. Here we estimate their quantitative importance, scaling by the lunar impact record. We give our results as an exogenous organic mass flux through time, to allow the influx to be readily determined for whatever epoch a particular model for the origin of life suggests is appropriate. Such an approach involves several uncertainties, about which we are explicit; we also argue, however, that these are no worse than many encountered in typical estimates of endogenous sources of prebiotic organics.

Attempts to estimate the impact environment of early Earth often begin with analytical fits to the lunar cratering record⁶⁻⁸, intended to minimize the model-dependence of the conclusions. In practice, however, this procedure faces numerous difficulties, which have often been disregarded. The oldest lunar province for which a radiometric date actually exists (the Apollo 16 and 17 uplands) is only 3.85–4.25 Gyr old; the ages of more heavily cratered provinces can at present only be estimated⁴. The entire interpretation of the heavy bombardment as representing exponentially decaying remnants of planetary formation is occasionally questioned by those favouring a lunar cataclysm²⁵. We have previously shown⁷ that different choices^{6,8} for the decay rates fitted to lunar cratering data^{4,5} can lead to substantially different conclusions about terrestrial mass influx during the heavy bombardment. More recently, we have argued that the more extreme of these choices can be excluded, as in contradiction with lunar and terrestrial geochemical data on meteoritic input²⁶. Here we employ a model lunar bombardment history that we have demonstrated is in good agreement with the geochemical constraints²⁶. Both the model and the constraints are summarized in Box 1.

Asteroid and comet impacts. The model developed in Box 1 for terrestrial mass accretion during the heavy bombardment is based on counts for lunar craters larger than 4 km in diameter. By equation (2), these correspond to impactors of masses greater than $\sim 10^{10}$ kg, or radii above ~ 100 m. What is the fate of organics in such objects incident on Earth? Some might catastrophically fragment ('airburst') while traversing the atmosphere; this case is treated below. We have demonstrated that such objects that do not airburst are insufficiently decelerated ('aerobraked') by a 1-bar terrestrial atmosphere for their organics to survive the heat of impact¹⁵. But some models suggest

that Earth may have had a dense, ~ 10 -bar CO_2 atmosphere some time before 3.8 Gyr ago^{29,30}. Under these conditions, comets with radii as large as ~ 100 m would have been sufficiently aerobraked (after substantial ablation during atmospheric passage) for most remaining organics to have survived impact with the terrestrial ocean¹⁵. Carbonaceous asteroids of similar size would not have been sufficiently aerobraked. Our earlier work on this topic¹⁵ considered a broad range of possible early terrestrial impact environments, because of uncertainties in lunar cratering data⁷. Since then (see Box 1), we claim to have reduced these uncertainties by appealing to geochemical constraints²⁶.

Following Delsemme's analysis of comet Halley data¹⁶, we take comets to be $\sim 14\%$ organic carbon by mass. The mass flux of cometary organic carbon surviving impact at time t is given by

$$\dot{m}(t) = \dot{m}(0)[\dot{n}(t)/\dot{n}(0)] = \dot{m}(0)f(t) \quad (8)$$

where $\dot{n}(t)$ is given by equation (7), and, if we use the results of Chyba *et al.*¹⁵ for cometary ablation, deceleration and resulting organic pyrolysis, $\dot{m}(0) = 6.6 \times 10^2 (\psi/0.1) \text{ kg yr}^{-1}$. Here ψ is the mass fraction of the ancient impacting flux in the relevant size range (here, radii ≤ 100 m) that was cometary. We review the data pertaining to the value of ψ elsewhere¹³ and find it inconclusive. Certain models^{31,32} of outer planet formation imply comet fluxes through the inner Solar System that require the bulk of the heavy bombardment to have been cometary. But disparities between the cratering records of the terrestrial planets (the planets nearest the Sun) and those of some outer planet satellites suggest that comets could not have been the main component of the heavy bombardment population³³. Admittedly, as long as only the lunar cratering record can be assigned absolute ages, such an objection remains unconfirmed. For now,

it is best to treat ψ as a free parameter. Our results, labelled 'comet impacts' in Fig. 1, take $\psi = 0.1$, and scale linearly, so the effect of a different choice for ψ is evident. More recent numerical modelling, extending earlier two-dimensional hydrodynamic impact models¹⁵ to a full three dimensions, suggests³⁴ the two-dimensional results used here may underestimate organic delivery from cometary impacts by a factor of ~ 3 .

Catastrophic airbursts. Photometric observations of Earth-crossing asteroids imply that $\sim \frac{1}{3}$ of the current asteroid flux at Earth is C-type³⁵. Of 17 terrestrial craters for which impactor type may be identified, only two seem to have had a carbonaceous chondritic composition³⁶. Similarly, carbonaceous chondrites constitute only ~ 5 – 6% of stony meteorite falls¹⁴, although they seem to represent over a third of 10^2 – 10^6 g Prairie Network fireballs (with another third seemingly having cometary origins)³⁷. It therefore appears that carbonaceous chondrites have material strengths so low that they are typically unable to survive atmospheric passage without breakup³⁸. Evidence from meteorite falls^{20,38} seems consistent with a compressive strength for carbonaceous chondrites of $\sigma_c \approx 10^6$ Pa (10 bars), orders of magnitude below typical rock strengths. In this case, the catastrophic fragmentation of the Revelstoke object²¹ may be a typical fate for carbonaceous impactors. Because of the increased surface-area-to-volume ratio of the resulting fragments, such airbursts might then provide an efficient mechanism for exogenous organics to reach Earth^{15,24}.

We estimate the magnitude of this source as follows. Melosh³⁹ has calculated the critical radius R_c of a meteoroid, greater than which—even should the object fragment—the fragments will have insufficient time to accelerate away from one another and follow independent trajectories before striking the ground. Only objects smaller than R_c are therefore able to deposit the bulk of their kinetic energy explosively in the atmosphere (that is, to

BOX 1 Impact statistics on the early Earth

Data⁴ for the cumulative surface density of lunar craters with diameter $> D$, as a function of surface age t in billions of years (Gyr), are well modelled by an equation of the form^{6,8}

$$N(t, D) = \alpha [t + \beta(e^{t/\tau} - 1)] (D/4,000 \text{ m})^{-1.8} \text{ km}^{-2} \quad (1)$$

We take $\tau = 144$ Myr, corresponding to a 100-Myr decay half-life for the impactor population, $\alpha = 3.5 \times 10^{-5}$, and $\beta = 2.3 \times 10^{-11}$ (refs 7, 26). Relying on the recent modelling of post-excavation crater collapse by McKinnon *et al.*²⁷, we have related²⁶ crater diameter D (in m) observed on the Moon to the mass m (in kg) and collision velocity v (in m s^{-1}) of an impactor by

$$m = 0.54 \gamma v^{-1.67} D_c^{0.44} D^{3.36} \quad (2)$$

where $D_c \approx 1.1 \times 10^4$ m is a transition diameter at which lunar craters change morphology from simple to complex forms²⁷, and $v = 1.2 \times 10^4 \text{ m s}^{-1}$ for typical impacts on the Moon. The constant $\gamma = 1.4 \times 10^3 \text{ kg s}^{-1.67} \text{ m}^{-2.13}$ depends on target and impactor densities, surface gravity, and impactor incidence angle. Equations (1) and (2) may be combined to give the number of objects of mass $> m$ that have struck the Moon as a function of time t :

$$n(> m, t) = 3.1 \times 10^8 [t + 2.3 \times 10^{-11}(e^{t/\tau} - 1)] m^{-b} \text{ kg}^b \quad (3)$$

where $b = 0.54$. The total mass, $M(t)$, incident on the Moon after some time t in impactors with masses in the range $m_{\min}$ to $m_{\max}$ is given by the integral:

$$M(t) = \int_{m_{\min}}^{m_{\max}} m [\partial n(> m, t) / \partial m] dm \quad (4)$$

which yields

$$M(t) = 3.7 \times 10^8 [t + 2.3 \times 10^{-11}(e^{t/\tau} - 1)] (m_{\max}^{1-b} - m_{\min}^{1-b}) \text{ kg}^b \quad (5)$$

To scale to the Earth, $M(t)$ must be multiplied by $\xi = 24$, the ratio of the terrestrial to the lunar gravitational cross sections at typical asteroid approach velocities²⁶.

Taking $m_{\max}$ to equal the mass of the lunar impactor that excavated the South Pole-Aitken basin ($D \approx 2,200$ km), or $\sim 1.4 \times 10^{19}$ kg by

equation (2), and $m_{\min}$ to be negligible, equation (5) yields a total mass of 1.0×10^{20} kg incident on the Moon subsequent to the solidification of the lunar crust ~ 4.4 Gyr ago. About half this mass would actually have been retained by the Moon. This result is in good agreement with the geochemical estimates of Sleep *et al.*¹⁰ of the meteoritic component mixed into the lunar crust, which yield $(0.4$ – $1.5) \times 10^{20}$ kg. Similarly, scaling equation (5) to Earth, and taking proper account of the statistical probability that the largest impactors incident on Earth were more massive than the largest incident on the Moon, gives an estimate of 1.5×10^{22} kg of material accumulated by Earth after 4.4 Gyr ago. This result is in good accord with geochemical estimates of post-core-formation meteoritic input. These estimates, based on chondritic abundances of highly siderophile elements in the terrestrial mantle, lie in the range $(1$ – $4) \times 10^{22}$ kg (refs 4, 26, 28). Certain proposed fits⁹ to other lunar cratering data sets⁵ seem to predict siderophile abundances two orders of magnitude in excess of those observed; such fits can probably be excluded as models for the early terrestrial impact environment. On the other hand, the close agreement of equation (5) with both the lunar cratering record and the available lunar and terrestrial geochemical constraints suggests that it provides a reasonable, albeit procrustean, model for post-core formation terrestrial mass accretion during the heavy bombardment.

Equation (5), multiplied by ξ gives the total mass incident on Earth in impactors within a certain size range, after some time t . A more useful quantity is the mass flux (kg yr^{-1}) at a particular time in the Earth's past. Combined with estimates of the organic mass fraction surviving delivery, this allows a quantitative comparison of these sources with photochemical or other *in situ* production rates of prebiotic organics as a function of time. To this end, we define $\dot{M}(t) = \xi [\partial M(t) / \partial t]$, the terrestrial mass flux from objects within a given mass range ($m_{\min}$ to $m_{\max}$) being accreted by Earth at a time t . From equation (5),

$$\dot{M}(t) = 8.9 f(t) (m_{\max}^{1-b} - m_{\min}^{1-b}) \text{ kg}^b \text{ yr}^{-1} \quad (6)$$

where $f(t) = (1 + 1.6 \times 10^{-10} e^{t/\tau})$. Our discussion will also require $\dot{n}(t) = \xi [\partial n(> m, t) / \partial t]$, the number flux of objects within a given mass range being accreted by Earth as a function of time. From equation (3),

$$\dot{n}(t) = 7.4 f(t) (m_{\max}^{-b} - m_{\min}^{-b}) \text{ kg}^b \text{ yr}^{-1} \quad (7)$$

airburst). For an incidence angle of 45° , the critical radii for icy and stony meteoroids in a 1-bar terrestrial atmosphere are 407 m and 235 m, respectively.

These results are consistent with the few relevant observations that are available. The Tunguska object is typically estimated^{19,20} to have had a mass of several billion (10^9) kilograms, corresponding to a comet ~ 100 m in radius. It seems to have exploded at an altitude of 6–9 km (ref. 19). The Lappajärvi crater, apparently the result of an impactor with a carbonaceous chondritic composition³⁶, required an object with a diameter of ~ 1 km; this impactor evidently did not airburst.

An approximate upper limit to organic delivery from airbursting asteroids and comets may therefore be obtained by taking $m_{\max}$ in equation (6) to be that appropriate to 235-m chondrites and 407-m comets. We take 10% of the impact object mass (for objects below ~ 0.5 km in radius) to be cometary, with the remainder split evenly between carbonaceous and non-carbonaceous objects (which average together to give a 1.3% organic carbon content¹⁴). Results are readily scaled to other assumptions about impactor compositions. Equation (6) then takes the form of equation (8), with $\dot{m}(0) = 3.6 \times 10^4 \text{ kg yr}^{-1}$, and is shown in Fig. 1. Clearly this value is an upper limit, as it ignores organic destruction in the airburst event itself (the Tunguska object exploded with an energy of $\sim 10^{16}$ J, or several megatons of high explosive²⁰) or in subsequent ablation of the resulting fragments. Moreover, not all objects with radii $< R_c$ will airburst. If the early terrestrial atmosphere were as thick as ~ 10 bar, larger objects could airburst³⁹, and the upper limit derived here would increase by a factor of about 30.

Interplanetary dust particles (IDPs). Anders¹⁴ has estimated the flux of intact organic matter reaching the contemporary Earth in IDPs and meteorites. The bulk of the mass in IDPs is in the $\sim 100\text{-}\mu\text{m}$ radius range, or, for a typical initial density⁴⁰ $\sim 1 \text{ g cm}^{-3}$, in particles with masses $\sim 10^{-5} \text{ g}$. The mass scaling in equation (6) leads one to expect most incident mass to lie in the largest particles, but in fact this scaling breaks down for particles with masses $\leq 10^2 \text{ g}$, probably because of dust production from larger bodies⁴¹. The global mass flux in particles below 10^2 g is observed^{42,43} to increase with decreasing size until it peaks at $\sim 10^{-5} \text{ g}$. This peak reaches a flux $\sim 10^5\text{--}10^6$ times higher than would be the case if the IDP mass spectrum did not deviate from power-law scaling.

Anders takes IDPs to be $\sim 10\%$ organic carbon by mass, a mean carbon content determined from 30 IDPs by X-ray analysis⁴⁴. He takes IDPs in the $10^{-12}\text{--}10^{-6} \text{ g}$ ($\sim 0.6\text{--}60 \mu\text{m}$ radius) range to be sufficiently gently decelerated during atmospheric entry to deliver their organics intact; his $60\text{-}\mu\text{m}$ upper limit is in good agreement with results of both theoretical modelling⁴⁰ and direct examination of IDPs⁴⁵. His lower limit is based on an estimate of the size below which IDP organics would be destroyed by ultraviolet photolysis; this choice could vary considerably without significant quantitative effects on the final result. It remains unclear whether the more fragile organic species would survive IDP atmospheric passage²⁴, especially at the high velocities appropriate to cometary-derived particles⁴⁰; there is probably a kind of deceleration-heating natural selection of the thermally most stable species.

Earth is currently accreting⁴¹ $\sim 3.2 \times 10^6 \text{ kg yr}^{-1}$ of $10^{-12}\text{--}10^{-6} \text{ g}$ IDPs, or $\dot{m}(0) \approx 3.2 \times 10^5 \text{ kg yr}^{-1}$ of intact organics¹⁴. It is unclear how we should scale this back in time through the heavy bombardment: would the mass influx peak at $\sim 10^{-5} \text{ g}$ have persisted? There are some suggestive, although less than compelling, data relevant to this question.

The current terrestrial mass influx from IDPs found by Hughes⁴² from meteor observations agrees well with accretion rates inferred from terrestrial and lunar data. As Table 1 shows, these data rely on a number of different sampling methods, and show a remarkably constant net IDP mass flux over the past ~ 3.6 Gyr, suggesting that the current IDP mass flux is at least not a recent anomaly. Most of these data, however, take the

form of an integral over the IDP size distribution, so cannot prove that the shape of the flux curve has remained unchanged with time. In any case, they do not extend further back than the end of the heavy bombardment.

Most IDPs currently collected in the stratosphere belong to one of three main classes; on the basis of particle heating and compositional data, these have been tentatively identified as corresponding to origins in main-belt asteroids, comets with perihelia outside 1.2 AU, and Earth-crossing asteroids or comets⁴⁶. The first two sources seem to be the most common. Main-belt asteroids seem unlikely to be a source representative of 'typical' heavy bombardment sources for IDPs, but Earth-crossing objects and comets with perihelia beyond 1.2 AU might well be.

Clearly we cannot say with confidence how to scale the IDP flux back beyond 3.6 Gyr ago. If IDPs are the product of cometary evaporation or of asteroid-asteroid collisions, one might expect their number to scale linearly, or as the square, respectively, of the number of such objects in the inner Solar System. Loss mechanisms, however, must also be considered. Whipple⁴⁷ has argued that, for dust whose loss rate is collisionally dominated, terrestrial accretion will scale as the square root of the cometary flux. But those IDPs actually able to contribute organics intact to Earth ($m \leq 10^{-6} \text{ g}$) have lifetimes dominated by Poynting-Robertson drag, not by collisions⁴³. The relative importance of these two loss mechanisms could have changed in an early Solar System in which the IDP flux was greater. Production of $\leq 10^{-6} \text{ g}$ particles (whose collisional production at present greatly exceeds their collisional destruction) would, however, also have increased as the square of the number of larger colliding particles. Relative production and destruction rates would have to be simultaneously modelled. To our knowledge, no such model currently exists.

Here we scale the IDP flux linearly with the lunar impact record, using equation (8) with $\dot{m}(0) \approx 3.2 \times 10^5 \text{ kg yr}^{-1}$ of organic carbon. A reasonable lower limit for the IDP flux on

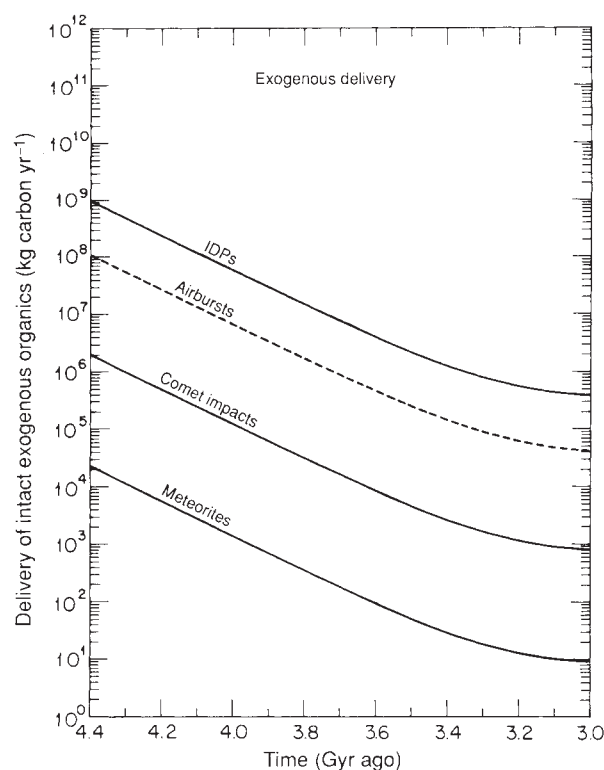


FIG. 1 Exogenous organic carbon delivered to the Earth as a function of time. Labels correspond to the appropriate sections of the text, where uncertainties are discussed. The dashed line indicates an upper bound.

Earth during the heavy bombardment is the current value. At 4 Gyr ago, this leads to an uncertainty of a factor of $\sim 10^2$.

Meteorites. Anders¹⁴ has estimated contemporary terrestrial accretion of intact organic matter from meteorites, finding 7.6 kg carbon yr^{-1} . Taking this value as $\dot{m}(0)$ in equation (8) gives the meteoritic organic carbon input through time, as shown in Fig. 1. Despite uncertainties, this source is clearly negligible compared with IDPs.

Accretion of interstellar dust. An additional source of exogenous organics on early Earth, independent of the heavy bombardment, would have been the terrestrial accretion of dust as the Solar System passed through interstellar clouds. Greenberg⁴⁸ has estimated that during its first 7×10^8 years, Earth should have passed through ~ 4 – 5 such clouds, accreting organic molecules during each $\sim 6 \times 10^5$ yr passage at a rate 10^6 – 10^7 kg yr^{-1} . Even during cloud passage, this source would have been one to two orders of magnitude smaller than that due to IDPs.

Shock synthesis of organic molecules

Gilvarry and Hochstim suggested that shock waves from meteoroids traversing the terrestrial atmosphere may have synthesized organics on early Earth⁴⁹. Bar-Nun *et al.* demonstrated that shock heating of reducing gas mixtures in the laboratory yielded amino acids in high yield¹⁷, although more recent results⁵⁰ suggest that the yields first reported¹⁷ for amino acid synthesis in $\text{CH}_4/\text{C}_2\text{H}_6/\text{NH}_3/\text{H}_2\text{O}$ atmospheres were overestimated by a factor of ~ 30 . Nevertheless, atmospheric shock synthesis may have been an important source of terrestrial organics^{17,18}. Estimates of both the early terrestrial impact environment and the coupling efficiency of an impactor's energy with the atmosphere may now be updated.

The mass of organics shock-synthesized in the atmosphere by an impactor is proportional to η , the organic synthesis efficiency (kilograms organic carbon produced per joule of shock energy): η is strongly dependent on atmospheric composition. Earlier models for a primordial reducing terrestrial atmosphere rich in CH_4 and NH_3 are now less favoured than that of a neutral atmosphere rich in CO_2 and N_2 . (Geochemical and photochemical arguments for this conclusion have been summarized in, for example, refs 15 and 29.) Nevertheless, both because this recent preference is not beyond doubt, and because many atmospheric compositions intermediate between the two extremes have been considered, we will treat two oxidation state end-members of a continuum of possible early-Earth atmospheres, one rich in CH_4 , the other rich in CO_2 .

A thermochemical model of shock synthesis in reducing $\text{CH}_4/\text{N}_2/\text{H}_2\text{O}$ atmospheres⁵¹ gives an HCN production efficiency $\sim 10^{17.5}$ molecules J^{-1} , in excellent agreement with laboratory results^{50,52} of $\sim 2 \times 10^{17}$ molecules J^{-1} . These experiments also show simultaneous formation of simple hydrocarbons, such as C_2H_2 and C_2H_4 , as well as carbon soot⁵². Although these results are temperature-dependent, typical organic C yields, ignoring soot, are ~ 3 times that of HCN alone⁵².

Because of the strength of the N_2 bond, we might expect efficiencies to increase in atmospheres where nitrogen is present as NH_3 , rather than N_2 . In fact, this effect is unimportant at the high shock temperatures appropriate here. Both high-power laser simulation, and theoretical high-temperature-equilibrium modelling, of shocks in NH_3/CH_4 atmospheres give HCN production yields⁵³ of $\sim 3 \times 10^{17}$ molecules J^{-1} . Hydrocarbons at least up to pentane (C_5H_{12}) are also produced. Using the experimental yields reported for all C-containing organic species gives a total efficiency for organic C production $\eta \approx 1.2 \times 10^{-8}$ kg J^{-1} for a reducing atmosphere⁵³, a result virtually identical with that found for N_2/CH_4 atmospheres^{50–52}. Efficiencies for HCN production in $\text{CO}_2/\text{N}_2/\text{H}_2\text{O}$ atmospheres, are $\sim 10^{7.5}$ times smaller⁵¹. In this case, formaldehyde (H_2CO) production is comparable¹⁸ to that of HCN, giving $\eta \approx 2.5 \times 10^{-16}$ kg J^{-1} for CO_2 atmospheres.

Shocks from meteors. Roughly 1.6×10^7 kg yr^{-1} strikes the Earth's atmosphere in meteors of mass 10^{-14} – 10^2 g (refs 14, 41). These objects lose 100% of their kinetic energy to the atmosphere. Assuming an average velocity of 15 km s^{-1} , about midway between those typical for asteroidal and cometary IDPs⁴⁰, meteors deposit 1.8×10^{15} J yr^{-1} into the present atmosphere. Some fraction, e_c , of this energy is converted into atmospheric shock waves. Pollack *et al.*^{54,55} have selected a value $e_c \approx 0.3$ for the efficiency of converting impactor kinetic energy into atmospheric shock heating, based on analyses of meteorite ablation in the contemporary atmosphere⁵⁶. With this value, total organic production in the atmosphere by meteors (Fig. 2) is given by $\dot{m}(t) = \eta(6 \times 10^{14} \text{ J yr}^{-1})f(t)$. Again, this result assumes that scaling with the lunar impact record is appropriate.

Shocks from airbursts. Objects as large as ~ 200 – 400 m in radius may airburst (see above), thereby coupling their entire kinetic energy to the atmosphere⁵⁷. Equation (6) can be used to put an upper limit on the importance of this effect. If we assume that all objects with radii ≤ 300 m will airburst, then Equation (6), for this choice of m_{max} and multiplied by $v^2/2$ (we take $v = 15 \text{ km s}^{-1}$, the median terrestrial collision velocity of Earth-crossing asteroids²⁶), gives $\dot{m}(t) = \eta(1.5 \times 10^{14} \text{ J yr}^{-1})f(t)$. This upper limit is shown in Fig. 2.

Shocks from post-impact vapour plumes. For large impactors, the total energy imparted to the atmosphere by the rapidly expanding post-impact plume is much greater than that lost by the impactor during atmospheric passage^{58,59}. Moreover, for a large impact, much of the atmosphere shocked by the incoming object will be shocked again by ejecta, largely pyrolysing products just synthesized during atmospheric passage⁶⁰. Meteors (see above) deliver about ten times as much net shock energy to the atmosphere than does the atmospheric passage of all impactors

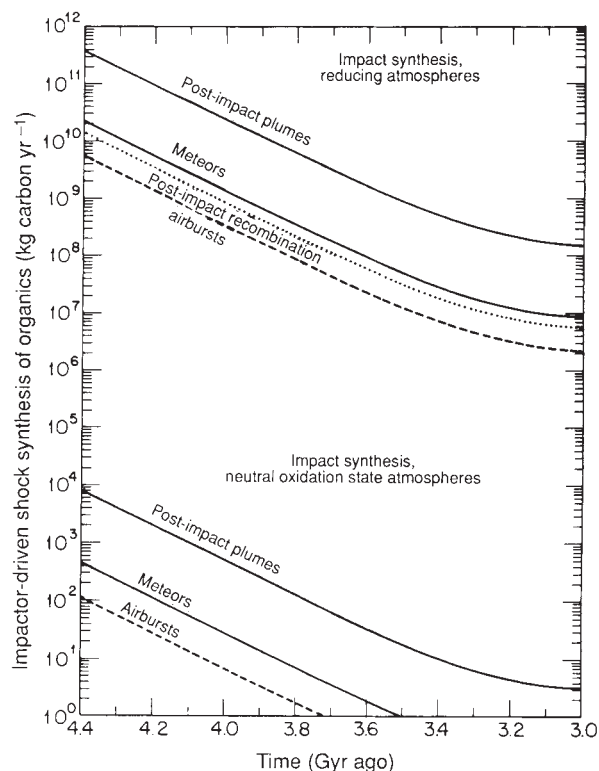


FIG. 2 Impactor-driven shock synthesis of organics in the terrestrial atmosphere as a function of time. Labels correspond to the appropriate sections of the text, where uncertainties are discussed. The upper curves are for CH_4 ($+\text{N}_2$ or $+\text{NH}_3$) $+\text{H}_2\text{O}$ reducing atmospheres, and the lower curves are for CO_2 $+\text{N}_2$ $+\text{H}_2\text{O}$ neutral oxidation state atmospheres. Dashed lines indicate upper bounds, and the dotted line denotes an estimate that is especially poorly understood, although perhaps largely independent of the oxidation state of the atmosphere.

TABLE 1 Estimates of terrestrial meteoric mass accretion

Sampling method	Effective sampling time	Mass influx (10^7 kg yr^{-1})
Atmospheric ⁸⁶	~1 yr	0.6–1.1
Satellite, radio, and visual observations ^{14,42,43}	past few decades	0.3–8.0*
Pleistocene abyssal clays ⁸⁷	past ~ 10^6 yr	3–9†‡
Cretaceous and Tertiary abyssal clays ⁴¹	33–67 Myr ago	1–5‡
Micrometeorite component in lunar soils ^{48,49}	past ~3.6 Gyr	1.8–2.8§

* Hughes' preferred value⁴² is $1.6 \times 10^7 \text{ kg yr}^{-1}$, with an uncertainty (adopted here) of 'at least half an order of magnitude'.

† As corrected by Kyte and Wasson⁴¹.

‡ After subtracting contribution from larger objects, following Kyte and Wasson⁴¹.

§ Scaled by a factor of 24 from the Moon to the larger gravitational cross section of Earth²⁶.

with masses $>10^2 \text{ g}$ combined⁶¹. Therefore we neglect atmospheric passage of large impactors, and concentrate on shock processing due to post-impact (plume) effects.

The simplest treatment of this problem would be first to multiply Equation (6) by $v^2/2$, to calculate the net kinetic energy flux at the terrestrial surface through time. Multiplying this result by η , and by ϵ , the fraction of an impactor's kinetic energy at surface impact returned to the atmosphere as shock energy, would then give an atmospheric organic production rate. We find $\dot{m}(t) = \eta\epsilon(6 \times 10^{17} \text{ J yr}^{-1})f(t)$. Typical published values⁵⁹ for ϵ are $\sim 1/8$.

This approach, however, ignores several important effects. Most important, the plumes arising from the largest impacts are poorly matched to the atmosphere; they rise far above the atmosphere and encounter only a small fraction of it⁶⁰. The result is that the shock processing yield per unit of impact energy decreases as the energy of the impact increases. (Failing to take such effects into account leads, following the procedure in the preceding paragraph, to the absurd conclusion that the largest objects (South Pole–Aitken size) striking the Earth would have shock-processed ~50% of an early reducing atmosphere.) Zahnle⁶⁰ has modelled these effects in detail for the case of impact production of nitric oxide in the present atmosphere. For an impact energy of $2 \times 10^{25} \text{ J}$, he finds a NO yield $\sim 1/40$ that calculated by Prinn and Fegley⁵⁹ with the assumption of a constant NO production yield (independent of impact energy). Kasting⁶² has integrated Zahnle's results over an impactor mass (hence, kinetic energy) distribution extending from 10^{13} to 10^{27} J , and finds an average NO production yield of 3.3×10^{14} molecule per joule of impact energy. (Kasting's choice of upper limit for impact energy is nearly identical to the one that we take for the South Pole–Aitken impactor; moreover, production efficiency depends only very weakly on the exact choice of exponent for the impactor mass distribution. We may therefore adopt his integration, an approximate treatment in any case, for our purposes here.) Prinn and Fegley⁵⁹ use a constant production yield of 2×10^{16} molecule per joule of energy imparted to the atmosphere. (Taking only $1/8$ of an impactor's kinetic energy to be so imparted, their production yield is really 2.5×10^{15} molecule per joule of impact energy.) In laboratory shock-tube experiments, η measures production per joule imparted to the atmosphere, so Kasting's result implies that η for shock production of organics should analogously be scaled down by a factor $(3.3 \times 10^{14})/(2 \times 10^{16}) \approx 1/60$ for production by impact plumes. We therefore have $\dot{m}(t) = \eta(1/60)(6 \times 10^{17} \text{ J yr}^{-1})f(t)$, as shown in Fig. 2.

The resulting timescale at 4.4 Gyr ago for complete shock processing of a 1-bar CH_4 atmosphere is $\sim 10^8 \text{ yr}$. A 1-bar CH_4 atmosphere on early Earth could therefore only have been

sustained if terrestrial volcanism or exogenous sources resupplied CH_4 to the atmosphere at a rate $\sim 2 \times 10^7 f(t) \text{ kg yr}^{-1}$, or $5 \times 10^{10} \text{ kg yr}^{-1}$ 4.4 Gyr ago. Terrestrial volcanoes are estimated⁶³ to release as much as $\sim 4 \times 10^{10} \text{ kg yr}^{-1}$ of carbon into the atmosphere, today mostly in the form of CO_2 . If carbon were released on a reducing primitive Earth as CH_4 , early terrestrial volcanism would need to have been no more intense than today to maintain an early reducing atmosphere against impactor shock-processing.

Finally, we note that the effects of erosion of the atmosphere because of especially energetic impacts are unlikely to have substantially altered the results found here. Only ~10% of asteroid, and ~50% of comet, collisions with the Earth occur at velocities high enough to cause erosion²⁶, effects small compared with other uncertainties in the problem (unless comets made up the bulk of heavy bombardment impactors).

Post-impact recombination. Several authors have suggested that organic molecules may have formed on early Earth by recombination from reducing mixtures resulting from the shock vaporization of bolides on impact^{64–66}. Others have suggested organic molecules might be similarly produced by impact shocks of terrestrial rocks^{64,67}. McKay *et al.* have attempted to simulate the former process using a thermochemical model⁶⁵. Their preliminary results show evidence for post-impact organic shock synthesis. But these simulations assumed as initial conditions a reducing mixture equivalent to the elemental composition of comets. In an actual impact, as much target material (oxygen-rich surface rocks) as impactor would be incorporated into the expanding and cooling vapour plume⁶, and background atmosphere, possibly CO_2 -rich, might also be entrained. The applicability of results for an initially reducing vapour mixture ($[\text{O}] < [\text{H}_2]$) is therefore unclear. Kasting⁶² argues that organic carbon in the hot rock vapour plume resulting from an impact would have been nearly entirely converted to CO.

At the same time, there are some relevant experimental data. Barak and Bar-Nun have demonstrated amino acid shock synthesis even when initial gas mixtures contain large quantities of H_2O and air⁶⁸. Mukhin *et al.* have stimulated shock vaporization and organic recombination using laser-pulse heating of a variety of terrestrial rocks and meteorite samples, and report a wide range of products of varying oxidation states⁶⁴. Unfortunately, their experiments vaporized only a fraction of the target, so organics may have been released from melted and heated target material, as well as synthesized in post-vaporization recombination. Experiments in which the entire target is vaporized would be informative.

In the Mukhin *et al.* experiment, most carbon was incorporated into CO and CO_2 , but typically several per cent went into hydrocarbons, and smaller amounts into HCN and aldehydes. The resulting ratio of $(\text{CO} + \text{CO}_2)$ to hydrocarbons was roughly constant, even as the initial wt% carbon of the sample varied over more than an order of magnitude; this seems to be consistent with post-shock recombination, rather than escape of unshocked organics. Hydrocarbon production (like that of CO and CO_2) scaled with the carbon content of the sample. We will consider the result of directly scaling these experiments up to impacts of comets and asteroids with the early Earth. First, we note that organic production from surface rocks is unimportant (~10%) compared with that for the carbon-rich impactors, so we henceforth ignore the former. Taking 10% of the impactor mass to be cometary (~17% C), of which 50% is blown out into space²⁶, and the remainder asteroidal (~1.3% C), and taking the Mukhin *et al.* result that ~4% of impactor carbon is incorporated into organics, we find post-impact recombination produced organic carbon at a rate $\dot{m}(t) = (4.6 \times 10^6 \text{ kg yr}^{-1})f(t)$ on the early Earth. To the extent that recombination occurs without significant entrainment of atmosphere, this result is independent of terrestrial atmosphere oxidation state. We emphasize the great uncertainties in the preceding calculation by displaying post-impact recombination results as a dotted line in Fig. 2.

TABLE 2 Inventory of main sources of prebiotic organics on Earth 4 Gyr ago

Source	Energy dissipation (J yr ⁻¹)	Production efficiency, reducing atmosphere (kg J ⁻¹)	Production efficiency, neutral* atmosphere (kg J ⁻¹)	Organic production, reducing atmosphere (kg yr ⁻¹)	Organic production, neutral* atmosphere (kg yr ⁻¹)
Lightning	1 (18)†	3 (-9)	3 (-11)	3 (9)	3 (7)
Coronal discharge	5 (17)	3 (-10)	3 (-12)	2 (8)	2 (6)
Ultraviolet light ($\lambda < 2,700 \text{ \AA}$)‡	1 (22)	2 (-11)	—	2 (11)	—
Ultraviolet light ($\lambda < 2,300 \text{ \AA}$)§	5 (21)	—	5 (-14)	—	3 (8)
Ultraviolet light ($\lambda < 2,000 \text{ \AA}$)	6 (20)	5 (-12)	—	3 (9)	—
Atmospheric shocks (meteors)	1 (17)	1 (-8)	3 (-16)	1 (9)	3 (1)
Atmospheric shocks (post-impact plumes)	1 (20)	2 (-10)¶	4 (-18)	2 (10)	4 (2)
IDPs	—	—	—	6 (7)	6 (7)
Totals	—	—	—	2 (11)	4 (8)

* $[\text{H}_2]/[\text{CO}_2] \approx 0.1$.† 1(18) should be read 1×10^{18} .‡ Appropriate to absorption by H_2S (ref. 79).§ Appropriate to absorption by CO_2 (ref. 74).|| Appropriate to absorption by H_2O in $\text{H}_2\text{O}/\text{CH}_4$ atmospheres (ref. 78).¶ Following Kasting⁶², averaged over Zahnle's⁶⁰ results for impactor energies from 10^{13} J to 10^{27} J.

Endogenous sources of prebiotic organics

At the time of the origins of life ~ 4 Gyr ago, IDPs may have been delivering $\sim 10^8 \text{ kg yr}^{-1}$ of organic carbon, regardless of the oxidation state of the terrestrial atmosphere. In the case of an early reducing atmosphere, shocks by post-impact plumes would have been producing organic carbon at $\sim 10^{10} \text{ kg yr}^{-1}$. How important would these heavy bombardment sources have been? We can put such rates into context by comparing them to the main endogenous sources of organics on the early Earth. Our list of possible terrestrial sources is not meant to be exhaustive. We do, however, explicitly cite those endogenous sources that have traditionally been considered to be the most important (Table 2). In the case of sources producing organics in the terrestrial atmosphere, Table 2 lists estimates of energy dissipation (J yr^{-1}), and experimentally derived values of production efficiency of organics (kg C J^{-1}) for reducing and neutral atmospheres. Organic production rates (kg yr^{-1}) are then derived.

As our standard 'neutral' or intermediate oxidation state atmosphere, we take one in which $[\text{H}_2]/[\text{CO}_2] \approx 0.1$. This choice is by no means the least reducing plausible neutral atmosphere. Kasting has recently considered an atmospheric H_2 mixture ratio of $\sim 10^{-3}$, at the upper end of the range used in many photochemical models for the early Earth, and based on H_2 outgassing rate near the upper limit of the flux likely to have been produced by early volcanoes or by photostimulated reduction of ferrous iron in the surface ocean⁶². (On the other hand, large impacts may have enhanced the atmospheric CO/CO_2 ratio⁶².) We list $[\text{H}_2]/[\text{CO}_2] \approx 0.1$ in Table 1, as it is that neutral atmosphere in which organic delivery by IDPs just begins to match endogenous production. It has been experimentally demonstrated^{69,70} that organic production rates drop precipitously by orders of magnitude as the ratio $[\text{H}_2]/[\text{CO}_2]$ falls below ~ 1 , so that lower $[\text{H}_2]/[\text{CO}_2]$ ratios will strongly favour IDPs as the dominant early source of organics.

In a review⁷¹ of contemporary data for terrestrial lightning and coronal energy dissipation, we recommended values of $1 \times 10^{18} \text{ J yr}^{-1}$ and $5 \times 10^{17} \text{ J yr}^{-1}$, respectively. (These values are factors of about 20 and 120 times smaller than those originally suggested by Miller and Urey⁷² in their compilation of sources of organics on the prebiotic Earth. The greatest uncertainty in these estimates, which is difficult to quantify, may be the extrapolation from contemporary rates of electrical energy discharge back to the primitive Earth.) We reviewed experimental work on organic production efficiencies for these sources, finding

coronal discharge efficiencies to be about an order of magnitude less than those for lightning⁷¹. Schlesinger and Miller have found⁷⁰ that organic yields drop by $\sim 10^2$ for H_2CO , HCN , and amino acids⁶⁹ as $[\text{H}_2]/[\text{CO}_2]$ drops from several (for which values organic production is roughly equal to that in CH_4 atmospheres) to ~ 0.1 . We have therefore scaled production efficiencies for the latter atmosphere in Table 2 down by 10^2 relative to the reducing case.

Apart from electrical energy sources, a principal source of energy for organic production on the early Earth was ultraviolet light from the Sun. In the past decade it has been recognized, from observations of young solar analogue stars, that the early ultraviolet luminosity of the Sun would have been greater than is the case today (despite lower net solar luminosity)^{73,74}. In Fig. 3, we use the results of Zahnle and Walker⁷³ for the evolution of the solar spectrum in 50- $\text{\AA}$ wavelength intervals (their Fig. 10) to derive the evolution of solar ultraviolet energy at the Earth as a function of time. Their model includes the evolution of important solar spectral lines; we follow it from a 10^8 -yr-old Sun (4.5 Gyr ago) until the present. Our values for the current (4.6-Gyr-old) solar ultraviolet fluxes are drawn from Heroux

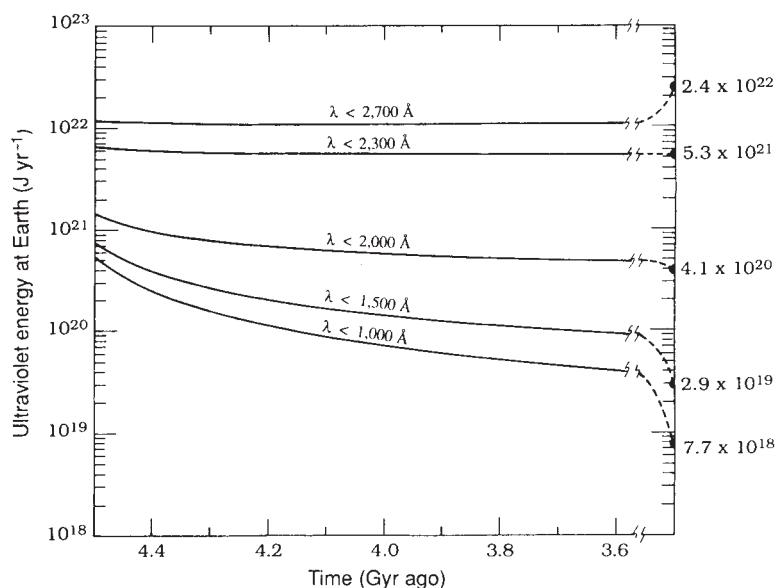


FIG. 3 Evolution of solar ultraviolet luminosity through time, beginning with a $\sim 10^8$ -Myr-old Sun at 4.5 Gyr ago, and extending to the present. Dashed lines beyond 3.6 Gyr ago indicate qualitative evolution to current values (labelled numerically).

and Hinteregger⁷⁵ for wavelengths below 2,000 Å, and from Arvesen *et al.*⁷⁶ from 2,000 Å to 2,700 Å.

Figure 3 shows the evolution of solar energy at the Earth for wavelengths below some selected value. The values shown are chosen as upper limits to the photodissociation continuum of gases that may have been of importance in the early atmosphere. For example, CH₄ is transparent at wavelengths above ~1,500 Å (ref. 77), H₂O above ~2,000 Å (ref. 78), CO₂ above ~2,300 Å (ref. 74), and H₂S above ~2,700 Å (ref. 79). Table 2 lists the net energy available on Earth below the latter three wavelengths at 4 Gyr ago (taking into account cosine incidence and diurnal effects).

In Table 2, the production efficiency for reducing (CH₄/H₂O) atmospheres (at wavelength $\lambda < 2,000$ Å) is taken from experimental work by Ferris and Chen⁷⁸, and that for atmospheres containing a long-wavelength ultraviolet photon acceptor generating superthermal hydrogen atoms, such as H₂S (or H₂CO), is taken from Sagan and Khare⁷⁹ and Hong *et al.*⁸⁰ (the latter authors demonstrate amino acid production using only CH₄ as a carbon source). These authors cite quantum yields for amino acids only; here we assume total organic carbon produced is ~10 times that incorporated into amino acids, a typical ratio^{69,70}. We assume that production efficiencies by ultraviolet acting on neutral atmospheres are ~100 times lower than those appropriate for CH₄ atmospheres, again following ref. 70. (This scaling agrees with photochemical modelling results⁸¹ for ultraviolet production of H₂CO in atmospheres where [H₂]/[CO₂] varies from ~3 to ~0.3.) For all these results it was assumed that production in a given candidate atmosphere was limited by the flux of ultraviolet light, not by the abundance of reactants—an assumption which may or may not have held for a given early Earth atmosphere.

An inventory of organics on early Earth

Our results for exogenous delivery, impact-shock synthesis and endogenous production of organic molecules 4 Gyr ago are summarized in Table 2. As indicated by Miller and Urey⁷², more energy is potentially available for organic synthesis from long-wavelength ultraviolet radiation than from any other source; and as indicated by Bar-Nun *et al.*¹⁷, shock synthesis has the highest specific efficiency. The total productivities listed in Table 2 show ~10¹¹ kg yr⁻¹ for reducing atmospheres 4 Gyr ago, and ~10⁸ kg yr⁻¹ for mainly neutral CO₂ atmospheres with 10% H₂. With lower [H₂], the endogenous productivities will be considerably less. Other atmospheres may have special mechanisms: for example, [CH₄] ~ 10⁻⁴ in an N₂ atmosphere will give⁸² ~10¹⁰ kg yr⁻¹ of HCN through short-ultraviolet photolysis. For both oxidation state end-member models, the heavy bombardment either generated or delivered quantities of organics roughly comparable to those from endogenous sources.

If all organic products were fully soluble in oceans of contemporary extent and depth, and had a mean lifetime of ~10⁷ yr against thermal degradation in mid-ocean vents or subducted plates, the steady-state organic abundance in the oceans 4 Gyr ago would have been ~10⁻³ g per g for a reducing atmosphere and 10⁻⁶ g per g for our intermediate oxidation state atmosphere. Although freeze-thaw and other concentration mechanisms doubtless existed, reducing atmospheres seem to be more favourable for the origin of life by three orders of magnitude or more; the entire ocean would then have been, almost literally, a 'dilute soup' of organic matter.

Qualitative as well as quantitative differences are likely. For example, IDPs are radiation-hardened from their stay in interplanetary space and may be rich in amorphous carbon and polycyclic aromatic hydrocarbons, which are less interesting for the origin of life than the amino acids and nucleotide bases produced by endogenous and impact processes. But a pathway from such hydrocarbons to amino acids has been proposed⁸³ (and disputed⁸⁴), and exogenous organics may be preferentially rich in amphiphilic vesicles, conceivably relevant to the origin

of the cell⁸⁵.

To summarize, which sources of prebiotic organics were quantitatively dominant depends strongly on the composition of the early terrestrial atmosphere. In the event of an early strongly reducing atmosphere, production by atmospheric shocks seems to have dominated that due to electrical discharges. Organic synthesis by ultraviolet light may, in turn, have dominated shock production, but only if a long-wavelength absorber such as H₂S were supplied to the atmosphere at a rate sufficient for synthesis to have been limited by ultraviolet flux, rather than by reactant abundance. In the apparently more likely case of an early terrestrial atmosphere of intermediate oxidation state, atmospheric shocks were probably of little importance for direct organic production. For [H₂]/[CO₂] ratios of ~0.1, net organic production was some three orders of magnitude lower than for reducing atmospheres, with delivery of intact exogenous organics in IDPs and ultraviolet production being the most important sources. At still lower [H₂]/[CO₂] ratios, IDPs may have been the dominant source of prebiotic organics on the early Earth. Endogenous, exogenous and impact-shock sources of organics could each have made a significant contribution to the origins of life. □

Christopher Chyba and Carl Sagan are at the Laboratory for Planetary Studies, Cornell University, Ithaca, New York 14853, USA. Christopher Chyba's present address is the Space Science Division, NASA Ames Research Center, Moffett Field, California 94035, USA.

- Schopf, J. W. & Walter, M. R. in *Earth's Earliest Biosphere* (ed. Schopf, J. W.) 214–239 (Princeton University Press, Princeton, 1983).
- Walter, M. R. in *Earth's Earliest Biosphere* (ed. Schopf, J. W.) 240–259 (Princeton University Press, Princeton, 1983).
- Schidlowski, M. *Nature* **333**, 313–318 (1988).
- Basaltic Volcanism Study Project, *Basaltic Volcanism on the Terrestrial Planets* (Pergamon, New York, 1981).
- Wilhelms, D. E. in *The Geology of the Terrestrial Planets*, NASA SP-469, 107–205 (1984).
- Melosh, H. J. & Vickery, A. M. *Nature* **338**, 487–489 (1989).
- Chyba, C. F. *Nature* **343**, 129–133 (1990).
- Maher, K. A. & Stevenson, D. J. *Nature* **331**, 612–614 (1988).
- Oberbeck, V. R. & Fogleman, G. *Nature* **339**, 434 (1989).
- Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. *Nature* **342**, 139–142 (1989).
- Oró, J. *Nature* **190**, 389–390 (1961).
- Anders, E. & Owen, T. *Science* **198**, 453–465 (1977).
- Chyba, C. F. *Nature* **330**, 632–635 (1987).
- Anders, E. *Nature* **342**, 255–257 (1989).
- Chyba, C. F., Thomas, P. J., Brookshaw, L. & Sagan, C. *Science* **249**, 366–373 (1990).
- Delsemme, A. H. in *Comets in the Post-Halley Era* (eds Newburn, R. L., Neugebauer, M. & Rahe, J.) 377–428 (Kluwer, Dordrecht, 1991).
- Bar-Nun, A., Bar-Nun, N., Bauer, S. H. & Sagan, C. *Science* **168**, 470–473 (1970).
- Fegley, B., Prinn, R. G., Hartman, H. & Watkins, G. H. *Nature* **319**, 305–308 (1986).
- Turco, R. P. *et al.* *Icarus* **50**, 1–52 (1982).
- Levin, B. Yu. & Bronshten, V. A. *Meteoritics* **21**, 199–215 (1986).
- Folinsbee, R. E., Douglas, J. A. V. & Maxwell, J. A. *Geochim. cosmochim. Acta* **31**, 1625–1635 (1967).
- Zhao, M. & Bada, J. L. *Nature* **339**, 463–465 (1989).
- Zahnle, K. & Grinspoon, D. *Nature* **348**, 157–160 (1990).
- Chyba, C. F. *Nature* **348**, 113–114 (1990).
- Ryder, G. *Eos* **71**, 313, 322–323 (1990).
- Chyba, C. F. *Icarus* **92**, 217–233 (1991).
- McKinnon, W. B., Chapman, C. R. & Housen, K. R. in *Uranus* (eds Bergstrahl, J. T., Miner, E. D. & Matthews, M. S.) (University of Arizona Press, Tucson, in the press).
- Dreibus, G. & Wänke, H. in *Origin and Evolution of Planetary and Satellite Atmospheres* (eds Atreya, S. K., Pollack, J. B. & Matthews, M. S.) 268–288 (University of Arizona Press, Tucson, 1989).
- Walker, J. C. G. *Origins Life* **16**, 117–127 (1986).
- Kasting, J. F. & Ackerman, T. P. *Science* **234**, 1383–1385 (1986).
- Fernández, J. A. & Ip, W.-H. *Icarus* **54**, 377–387 (1983).
- Shoemaker, E. M. & Wolfe, R. F. *Proc. Lunar Planet. Sci. Conf.* **15**, 780–781 (1984).
- Strom, R. G. *Icarus* **70**, 517–535 (1987).
- Thomas, P. J., Brookshaw, L., Chyba, C. F. & Sagan, C. *Eos* **71**, 1429 (1990).
- Shoemaker, E. M., Wolfe, R. F. & Shoemaker, C. S. in *Global Catastrophes in Earth History* (eds Sharpton, V. L. & Ward, P. D.) ESA SP-247, 155–170 (1990).
- Grieve, R. A. F. *Geol. Soc. Am. Spec. Pap.* **190**, 25–37 (1982).
- Ceplecha, Z. in *Comets, Asteroids, and Meteorites* (ed. Delsemme, A.) 143–152 (University of Toledo, Toledo, 1977).
- Baldwin, B. & Sheaffer, Y. J. *geophys. Res.* **76**, 4653–4668 (1971).
- Melosh, H. J. in *Multi-ring Basins, Proc. Lunar Planet. Sci.* (eds Schultz, P. H. & Merrill, R. B.) 29–35 (Pergamon, New York, 1981).
- Love, S. G. & Brownlee, D. E. *Icarus* **89**, 26–43 (1991).
- Kyte, F. T. & Wesson, J. T. *Science* **232**, 1225–1229 (1986).
- Hughes, D. W. in *Cosmic Dust* (ed. McDonnell, J. A. M.), 123–185 (J Wiley, New York, 1978).
- Grün, E., Zook, H. A., Fechtig, H. & Giese, R. H. *Icarus* **62**, 244–272 (1985).
- Schramm, L. S., Brownlee, D. E. & Wheelock, M. M. *Meteoritics* **24**, 99–112 (1989).
- Brownlee, D. E. *Annu. Rev. Earth planet. Sci.* **13**, 189–215 (1985).
- Sanford, S. A. & Bradley, J. P. *Icarus* **82**, 146–166 (1989).
- Whipple, F. L. *Mém. Soc. R. Liège Série 9*, 101–111 (1976).
- Greenberg, J. M. in *Comets and the Origin of Life* (ed. Ponnamperna, C.), 111–127 (Reidel, Dordrecht, 1981).
- Gilvarry, J. J. & Hochstim, A. R. *Nature* **197**, 624–625 (1963).

50. Bar-Nun, A. & Shaviv, A. *Icarus* **24**, 197–210 (1975).
51. Chameides, W. L. & Walker, J. C. G. *Origins Life* **11**, 291–302 (1981).
52. Rao, V. V., Mackay, D. & Trass, O. *Can. J. chem. Eng.* **45**, 61–66 (1967).
53. Scattergood et al. *Icarus* **81**, 413–428 (1989).
54. McKay, C. P., Scattergood, T. W., Pollack, J. B., Borucki, W. J. & Van Ghysseghem, H. T. *Nature* **332**, 520–522 (1988).
55. Pollack, J. B., Podolak, M., Bodenheimer, P. & Christofferson, B. *Icarus* **67**, 409–443 (1986).
56. Bronshten, V. A. *Physics of Meteoric Phenomena* (Reidel, Dordrecht, The Netherlands, 1983).
57. Lewis, J. S., Watkins, G. H., Hartman, H. & Prinn, R. G. *Geol. Soc. Am. Spec. Pap.* **190**, 215–221 (1982).
58. O'Keefe, J. D. & Ahrens, T. J. *Geol. Soc. Am. Spec. Pap.* **190**, 103–120 (1982).
59. Prinn, R. G. & Fegley, B. *Earth planet. Sci. Lett.* **83**, 1–15 (1987).
60. Zahnle, K. J. in *Global Catastrophes in Earth History* (eds Sharpton, V. L. & Ward, P. D.), GSA SP-247, 271–288 (1990).
61. Chyba, C. F. thesis, Cornell Univ. (1991).
62. Kasting, J. F. *Origins Life* **20**, 199–231 (1990).
63. Walker, J. C. G. *Evolution of the Atmosphere* (Macmillan, New York, 1977).
64. Mukhin, L. M., Gerasimov, M. V. & Safonova, E. N. *Nature* **340**, 46–48 (1989).
65. McKay, C. P., Borucki, W. R., Kujiro, D. R. & Church, F. *Lunar planet. Sci. Conf. XX* 671–672 (1989).
66. Oberbeck, V. R., McKay, C. P., Scattergood, T. W., Carle, G. C. & Valentin, J. R. *Origins Life* **19**, 39–55 (1989).
67. Freund, F., Dickinson, J. T., Becker, C. H., Freund, M. M. & Chang, S. *Origins Life* **16**, 291–292 (1986).
68. Barak, I. & Bar-Nun, A. *Origins Life* **6**, 483–506 (1975).
69. Schlesinger, G. & Miller, S. L. *J. molec. Evol.* **19**, 376–382 (1983).
70. Schlesinger, G. & Miller, S. L. *J. molec. Evol.* **19**, 383–390 (1983).
71. Chyba, C. & Sagan, C. *Origins Life* **21**, 3–17 (1991).
72. Miller, S. L. & Urey, H. C. *Science* **130**, 245–251 (1959).
73. Zahnle, K. J. & Walker, J. C. G. *Rev. Geophys. Space Phys.* **20**, 280–292 (1982).
74. Canuto, V. M., Levine, J. S., Augustsson, T. R. & Imhoff, C. L. *Nature* **296**, 816–820 (1982).
75. Heroux, L. & Hinteregger, H. E. *J. geophys. Res.* **83**, 5305–5308 (1978).
76. Arvesen, J. C., Griffin, R. N. & Pearson, B. D. *Appl. Opt.* **8**, 2215–2232 (1969).
77. Mount, G. H., Warden, E. S. & Moos, H. W. *Astrophys. J.* **214**, L47–49 (1977).
78. Ferris, J. P. & Chen, C. T. *J. Am. chem. Soc.* **97**, 2962–2967 (1975).
79. Sagan, C. & Khare, B. N. *Science* **173**, 417–420 (1971).
80. Hong, K.-Y., Hong, J.-H. & Becker, R. S. *Science* **184**, 984–987 (1974).
81. Pinto, J. P., Galdstone, G. R. & Yung, Y. L. *Science* **210**, 183–185 (1980).
82. Zahnle, K. J. *J. geophys. Res.* **91**, 2819–2834 (1986).
83. Shock, E. O. & Schulte, M. D. *Nature* **343**, 728–731 (1990).
84. Miller, S. L. & Bada, J. L. *Nature* **350**, 388–389 (1991).
85. Deamer, D. W. *Nature* **317**, 792–794 (1985).
86. Tuncel, G. & Zoller, W. H. *Nature* **329**, 703–705 (1987).
87. Barker, J. L. & Anders, E. *Geochim. cosmochim. Acta* **32**, 627–645 (1968).
88. Anders, E., Ganapathy, R., Krähenbühl, U. & Morgan, J. W. *Moon* **8**, 3–24 (1973).
89. Dohnanyi, J. S. *Science* **173**, 558 (1971).

ACKNOWLEDGEMENTS. We thank P. J. Thomas and L. Brookshaw for discussions, and K. Zahnle and H. J. Melosh for reviews of an earlier version of this manuscript. This work was supported by the Kenneth T. and Eileen L. Norris Foundation and NASA.

ARTICLES

Experimental constraints on strong-field relativistic gravity

J. H. Taylor*, A. Wolszczan†, Thibault Damour‡§ & J. M. Weisberg||

* Joseph Henry Laboratories and Physics Department, Princeton University, Princeton, New Jersey 08544, USA

† Arecibo Observatory, National Astronomy and Ionosphere Center, PO Box 995, Arecibo, Puerto Rico 00613, USA

‡ Institut des Hautes Etudes Scientifiques, 91440 Bures sur Yvette, France

|| Physics and Astronomy Department, Carleton College, Northfield, Minnesota 55057 USA

Experiments in our Solar System can test relativistic gravity only in the weak-field limit, but systems containing pulsars necessarily involve the effects of strong-field gravity. Timing observations of three binary pulsars yield tight constraints on the nature of gravity in the strong-field regime, allowing the measurement of velocity-dependent and nonlinear phenomena separately from the effects of gravitational radiation. General relativity passes these new experimental tests with complete success.

ALL tests of general relativity within the Solar System^{1,2}, including perihelion advances of planetary orbits, the bending and delay of light rays passing near the Sun, and limits on the violation of the strong equivalence principle in the Earth–Moon system, have examined the gravitational interaction only in connection with weakly self-gravitating objects (for example, for the Sun, $GM_{\odot}/c^2 R_{\odot} \approx 2 \times 10^{-6}$). The measured relativistic effects are therefore but small perturbations to newtonian expectations, usually very difficult to measure even with modern technologies. More significantly, even when the experimental accuracy is high these tests have an important qualitative weakness: they say nothing about how the ‘correct’ theory of gravity might behave when gravitational forces are very strong, such as near a neutron star or black hole. Pulsars in gravitationally bound binary orbits provide nearly ideal laboratories for the testing of strong-field gravity: they have surface gravitational potentials $GM/c^2 R \approx 0.2$; they move with mildly relativistic velocities ($v/c \approx 10^{-3}$) through a repetitive cycle well suited to

experimental averaging techniques; and they emit periodic pulses of radio noise, detectable over interstellar distances, in some cases as stable as the ticks of an atomic clock³.

Here we make use of 10 years of high-quality timing observations⁴ of the binary pulsar PSR1913+16, and 1 year of similar data⁵ for the newly discovered system PSR1534+12, to obtain experimental constraints on gravity in the strong-field regime. We use a methodology recently advocated by two of us (T.D. & J.H.T.; ref. 6) and we combine the new results with a published limit⁷ based on the binary pulsar PSR1855+09. The results are interpreted in terms of a parametrized family of tensor–biscalar theories containing general relativity as a special case⁸. Our experimental constraints impose significant limits on how far the ‘correct’ theory of gravity can deviate from general relativity in the strong-field, non-quantum regime.

A related use of binary pulsar data for establishing the existence of gravitational radiation is by now well known^{9–13}. This work is based on observations of times of arrival (TOAs) of pulses from PSR1913+16, which have been accumulating over more than 17 years. The TOAs are analysed by comparing them with predictions of a relativistic timing model^{14–17}, providing in the process accurate measurements of five ‘keplerian’ parameters of the orbit and three ‘post-keplerian’ ones: the advance of periastron $\dot{\omega}$, a time-dilation and gravitational redshift parameter γ , and the secular change of the orbital period, $\dot{P}_b$. Gravitational radiation is expected to remove energy and angular momentum from the orbit, causing the period to decrease. A quantitative test for gravitational radiation can therefore be carried out by comparing the observed $\dot{P}_b$ with the value predicted by general relativity (or another theory to be evaluated), given the measured values of $\dot{\omega}$, γ and the keplerian parameters.

General relativity now passes this $\dot{\omega} - \gamma - \dot{P}_b$ test for PSR1913+16 with an accuracy better than 0.5% (see further details below). But Damour and Esposito-Farèse⁸ have recently

§ Also at Département d’Astrophysique Relativiste et de Cosmologie, Observatoire de Paris, Centre National de la Recherche Scientifique, 92195 Meudon Cedex, France.

constructed examples of other theories that can also pass the test, as well as all known tests of gravity in the weak-field limit, while differing substantially from general relativity in the strong-field regime. They call attention to an inherent complication of the $\dot{\omega} - \gamma - \dot{P}_b$ test: it necessarily mixes quasi-stationary strong-field effects (related to $\dot{\omega}$ and γ) with radiative ones (related to $\dot{P}_b$). If a theory fails the test, one gains no hints as to whether its problems lie in the (quasi-stationary) strong field or the radiative regime. Alternatively, an invalid theory might pass the mixed test even though it would fail others based separately on radiative or strong-field effects. Damour and Taylor have addressed these issues and described in detail the 'parametrized post-keplerian' (PPK) formalism, an expanded phenomenological framework for extracting the maximum theoretical content from binary pulsar data⁶. Here we apply the PPK approach to observations of PSR1913+16 and PSR1534+12, with our main emphasis on obtaining significant experimental constraints on the nonradiative strong-field regime of gravitation theories.

Observations

All of the relevant observations have been made with the Arecibo Observatory's uniquely sensitive 305-m radio telescope, at frequencies near 430 and 1,400 MHz. PSR1913+16 has been observed regularly since 1974, mostly with the help of special-purpose data-acquisition equipment developed at the universities of Massachusetts and Princeton^{10,12}. A series of particularly high-quality observations at 1,400 MHz began in 1981, and the most recent observing session was held in December 1990. A total of 4,926 TOAs are now available for this pulsar, each based on ~ 5 min of data. Most of the 3,713 measurements made since 1981 have uncertainties around 15 μ s, and these make up the data set analysed here.

PSR1534+12 was first detected in February 1990, and timing measurements have been made at least several times per month since the discovery was confirmed in July of that year⁵. The pulsar is stronger than PSR1913+16 and has a sharper, narrower pulse, so its timing measurements are considerably more precise. More than 1,300 TOAs have been acquired for this pulsar at 430 MHz, using the same Princeton Mark III timing equipment as used for PSR1913+16. Typical estimated uncertainties (random plus systematic) for these observations are $\sim 6 \mu$ s for 2.5-min integrations. In addition, starting in October 1990, 406 TOAs have been obtained in the same way at frequencies near 1,400 MHz. These measurements are of substantially higher quality, with uncertainties of 3.5 μ s for 5-min integrations, and they show no evidence of significant systematic errors. Our results are based almost entirely on the 1,400-MHz data set. Further details of the timing and other observations of PSR1534+12 will be given elsewhere (A. Phillips, A.W., Z. Arzoumanian and J.H.T.; A.W. and J.H.T.; manuscripts in preparation).

A feeling for the very high precision of the data available for this experiment can be gained from Fig. 1. The smooth curves *a* and *b* illustrate the time delays caused by orbital motion of PSR1913+16 and PSR1534+12, respectively; the ordinate scales for the delays are given in seconds. The plotted points with error bars represent residual differences between the observed TOAs and the relativistic timing model, averaged into 20 equally spaced bins of orbital phase. For the residuals the vertical scale has been magnified 10^6 times, so the numbers are read in microseconds. In this sense the basic signal-to-noise ratio of the experiment is evidently about a million, and it is not surprising that timing effects of order (v^2/c^2) should prove accessible.

Phenomenological analysis

Most of the high-quality data on PSR1913+16 have already been analysed using the PPK approach, and full solutions and details of the methods have been published^{12,13}. An updated solution based on TOAs obtained through the end of 1990 was

quoted in ref. 6, and we now provide further details and fitted parameters. The PPK analysis involves fitting measured TOAs to the phenomenological timing model of ref. 17. The model is based on a treatment of the relativistic two-body problem by the same authors¹⁸, in which all timing effects up to order (v^2/c^2) are described in a manner common to a wide variety of possible gravitation theories. It includes five keplerian and eight post-keplerian orbital parameters; in principle all of them can be determined from timing data by standard optimization techniques⁶ (see also ref. 12).

Table 1 lists the best-fitting orbital parameters now obtained by the PPK method for PSR1534+12 and PSR1913+16. The keplerian parameters have been determined for both pulsars with precisions of seven or more significant digits; they include the orbital period P_b , eccentricity e , projected semimajor axis $x \equiv a_1 \sin i/c$, time of periastron passage T_0 and longitude of periastron ω . (Here a_1 is the semimajor axis of the pulsar orbit, i is the angle between the orbital angular momentum and the line of sight from Earth to pulsar, and c is the speed of light.) The post-keplerian parameters $\dot{\omega}$ and γ have also been determined with good accuracy for both systems, as has $\dot{P}_b$ for PSR1913+16. We call attention to the fact that detailed comparison of the observed $\dot{P}_b$ with theoretical predictions must allow for a small contribution due to kinematical effects in our Galaxy, amounting to $\Delta\dot{P}_b = (-0.017 \pm 0.005) \times 10^{-12}$ for PSR1913+16 (ref. 13). After this correction has been applied, the ratio of the observed $\dot{P}_b$ to the predicted effect for gravitational-radiation damping in general relativity is 1.0023 ± 0.0047 . The PPK parameters r and s in Table 1 are the two parameters entering the Shapiro time delay¹⁹ caused by the gravitational

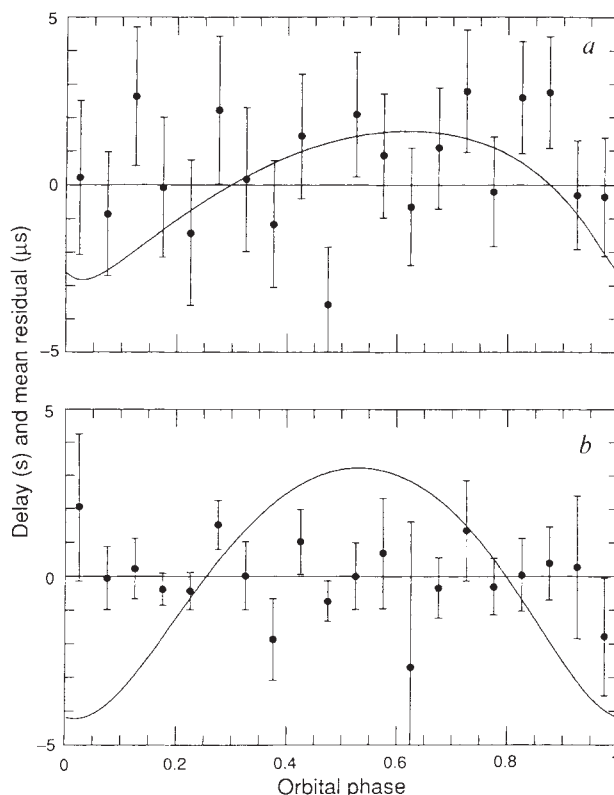


FIG. 1 Smooth curves represent time delays (in seconds) caused by the orbital motion of PSR1913+16 (*a*) and PSR1534+12 (*b*). Observations fit the theoretical models so well that if plotted on the same scale, the measurements would fall 'exactly' on the curves and their error bars would be shorter than the line thickness. To provide an indicator of data quality, we therefore plot the mean post-fit residuals (averaged into 20 equal bins of orbital phase) on a vertical scale magnified 10^6 times. Irregular sampling of orbital phases causes the variation in sizes of error bars for the PSR1534+12 data.

TABLE 1 Measured orbital parameters of two binary pulsar systems

	PSR1534 + 12	PSR1913 + 16
Keplerian parameters		
Orbital period P_b (s)	36,351.70270(3)	27,906.9807807(9)
Eccentricity, e	0.2736779(6)	0.6171309(6)
Projected semimajor axis, x (s)	3.729468(9)	2.341759(3)
Time of periastron, T_0 (MJD)	48,262.8434966(2)	46,443.99588321(5)
Longitude of periastron, ω ($^\circ$)	264.9721(16)	226.57531(9)
Post-keplerian parameters		
Advance of periastron, $\dot{\omega}$ ($^\circ \text{ yr}^{-1}$)	1.7560(3)	4.226628(18)
Time dilation, γ (ms)	2.05(11)	4.294(3)
Orbital period derivative, $\dot{P}_b$ (10^{-12})	-0.1(6)	-2.425(10)
Range of Shapiro delay, r (μs)	6.2(1.3)	(see Fig. 2)
Shape of Shapiro delay, $s = \sin i$	0.986(7)	(see Fig. 2)

Figures in parentheses represent uncertainties in the last quoted digit.

influence of the companion. More precisely, r measures the overall amplitude or 'range' of this delay, and s (numerically equal to $\sin i$) determines its 'shape'.

The parameters r and s are not separately measurable with great accuracy for PSR1913 + 16, although sizeable regions of their parameter space can be excluded with high confidence¹². The acceptable region of the (r, s) plane can be determined by calculating values of the goodness-of-fit statistic χ^2 over a wide range of fixed values of r and s . Level contours corresponding to $\Delta\chi^2(r, s) = 1.0, 2.3$ and 6.2 (where $\Delta\chi^2$ is the increment in χ^2 above its global minimum) are plotted in Fig. 2, labelled by the corresponding confidence levels, 39%, 68% and 95%. (A similar illustration and further details of the procedure used to calculate the contours may be found in ref. 12.) The filled circle at $r = 6.83 \mu\text{s}$, $s = 0.734$ corresponds to the parameter values expec-

ted for this pulsar according to general relativity (see equations (16), (17), (20) and (21) of ref. 12; see also ref. 17 for more details). With the present data on PSR1913 + 16, all values of $s \geq 0.9$ can be excluded with high confidence unless r is improbably small. The remaining parameter space still leaves wide latitude for strong-field gravitational effects quite different from those in general relativity, so this test of gravity by itself is not very restrictive.

The orbit of PSR1534 + 12 is more nearly 'edge-on' to the line of sight than that of PSR1913 + 16. This fact, together with the more accurate TOAs obtained for PSR1534 + 12, makes it possible to measure r and s separately for this pulsar, in addition to $\dot{\omega}$ and γ . The expected parameter values in general relativity, given the measured $\dot{\omega}$ and γ , are $r = 6.58 \mu\text{s}$ and $s = 0.982$. These values are denoted by a filled triangle in Fig. 2, and contours surrounding the triangle define the experimentally allowed range at the 68% and 95% confidence levels. The most probable values of r and s are listed in Table 1, along with the other orbital parameters. The data span for this pulsar is currently too short to permit a measurement of $\dot{P}_b$. But the upper limit in Table 1 (obtained from a fit including the 430 MHz as well as the 1,400 MHz data) is already close to the expected general relativistic value, roughly -1.925×10^{-13} . One may therefore expect that a reliable measurement of $\dot{P}_b$ should be possible within a few years, yielding an $\dot{\omega} - \gamma - \dot{P}_b$ test analogous to the one available for PSR1913 + 16.

We should emphasize that the measurement of four non-radiative PPK parameters for PSR1534 + 12, namely $\dot{\omega}$, γ , r and s , already gives access to two new, independent tests of quasi-stationary strong-field gravity.

Alternatives to general relativity

Following the guidelines in ref. 6, we now shift attention from a purely phenomenological to a theory-dependent analysis of the data. This will help to elucidate the physical meaning of the measured phenomenological parameters, transforming them into explicit quantitative limits on gravitation theories in the strong-field regime. It will also permit a useful combination of the qualitatively different constraints obtained from PSR1913 +

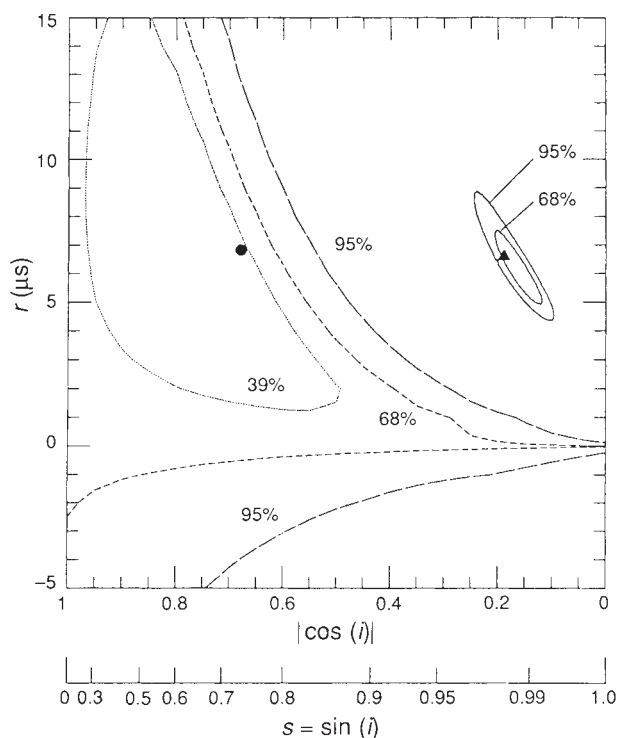


FIG. 2 Regions of the (r, s) plane consistent with timing observations of pulsars 1913 + 16 and 1534 + 12. Long contours extending across much of the panel represent 39%, 68% and 95% confidence regions for the r and s parameters of PSR1913 + 16. The values consistent with general relativity are indicated by a filled circle at $r = 6.83 \mu\text{s}$, $s = 0.734$. Closed contours surrounding the triangle at the right represent 68% and 95% confidence intervals for r and s in the PSR1534 + 12 system. The triangle marks the general relativistic prediction at $r = 6.58 \mu\text{s}$, $s = 0.982$.

16, PSR1534+12 and a third pulsar, PSR1855+09, to obtain the strongest possible overall limits. To do this we start from the very general class of tensor-multiscalar theories studied recently in ref. 8. These theories arise naturally in attempts to quantize gravity or to unify it with the other fundamental interactions. They are also the most natural generalizations of the well-known tensor-scalar theory of Jordan, Fierz, Brans and Dicke. Among this general class, we shall follow ref. 8 in concentrating our attention on a two-parameter subclass of tensor-biscalar theories, denoted $T(\beta', \beta'')$. Theories in this subclass have two scalar fields plus the usual tensor, $g_{\mu\nu}$, mediating the effective gravitational interaction, and they tend smoothly to general relativity when both parameters β' and β'' tend to zero. Our goal will be to place experimental limits on these two adjustable parameters, following the theoretical framework described in ref. 6.

Let us recall that in the familiar weak-field conditions a set of 'parametrized post-newtonian' (PPN) parameters $\beta, \gamma, \dots$, characterize possible weak-field deviations from general relativity: γ characterizes velocity-dependent or 'magnetic' effects, β the nonlinear gravitational effects. (Beware of the potentially confusing notation used here for historical continuity: the PPN γ , a dimensionless number of order unity, bears no direct relationship to the PPK measurable of the same name, as listed in Table 1.) The theories $T(\beta', \beta'')$ all have the same post-newtonian limit as general relativity; that is, their PPN parameters are exactly $\beta - 1 = \gamma - 1 = \dots = 0$ (so that they pass all existing Solar-System tests), but they deviate from Einstein's theory when describing systems comprising strongly self-gravitating bodies. To illustrate how the two numbers β' and β'' parametrize possible strong-field deviations from general relativity, let us consider the lagrangian describing, within the theory $T(\beta', \beta'')$, the relative motion of a system of two strongly self-gravitating bodies of inertial mass m_1 and m_2 . This lagrangian contains several interaction terms: a newtonian-like $1/R$ potential in which there appears an effective gravitational constant $\tilde{G}$, some velocity-dependent terms describable in terms of an effective PPN γ -parameter, say $\tilde{\gamma}$, and a nonlinear relativistic $1/R^2$ potential which can be associated with an effective PPN β -parameter, say $\tilde{\beta}$. These effective parameters depend on how strongly self-gravitating are the two bodies, as measured by the 'compactness' defined as twice the fractional gravitational binding energy, $c_A = 2|E_{\text{grav}}^{\text{body } A}|/m_A c^2$, for $A=1, 2$. The effective parameters entering the lagrangian have the form

$$\begin{aligned}\tilde{G} &= G_{\text{Newton}}[1 + \tfrac{1}{2}\beta'(c_1^2 + c_2^2)] \\ \tilde{\gamma} &= 1 - \frac{\beta'(c_1^2 + c_2^2)}{1 + \tfrac{1}{2}\beta'(c_1^2 + c_2^2)} \\ \tilde{\beta} &= 1 + \frac{\beta''(X_1 c_2^2 + X_2 c_1^2) + \beta' P_1(c_1, c_2) + \beta'^2 P_3(c_1, c_2)}{4[1 + \tfrac{1}{2}\beta'(c_1^2 + c_2^2)]^2}\end{aligned}$$

where $X_1 \equiv m_1/(m_1 + m_2) \equiv 1 - X_2$, and where P_1 and P_3 denote some rather complicated polynomial expressions which start at order 1 and 3, respectively, in the compactness parameters c_1, c_2 . (To simplify the reading, we have here absorbed a factor $B = 1.026 \dots$ into β' and β'' ; see refs 6, 8 for details, including the precise definition of the compactness parameters.) The equations above make it clear that constraining the values of $|\beta'|$ and $|\beta''|$ means constraining the possible influence of strong-field effect on the gravitational dynamics of bodies. In general relativity, both parameters β' and β'' take the value zero, and therefore $\tilde{G} \equiv G_{\text{Newton}}$, $\tilde{\gamma} - 1 \equiv \tilde{\beta} - 1 \equiv 0$.

Numerical values of c_1 and c_2 depend on the choice of an equation of state for describing the internal structure of a neutron star. Damour and Esposito-Farèse⁸ have made simple fits to four representative neutron-star models, and we have adopted the median value from their fits, namely a compactness given by $c_A = 0.21 m_A/M_\odot$. Changes in the experimentally allowed regions of the $\beta' - \beta''$ plane (see Fig. 3 below) when using more

extreme (very stiff, or very soft) equations of state have been studied in ref. 8; for a related example of how the choice of a neutron star model affects observationally deduced bounds on theoretical parameters, see ref. 20.

Constraints on parameters β', β''

A first constraint on the strong-field parameters β' and β'' can be obtained immediately, in favourable circumstances, when at least two phenomenological (PPK) parameters have been measured for a pulsar. As emphasized in ref. 6, each of the measured post-keplerian parameters defines, when interpreted within the framework of a specific theory of gravity, a curve relating the inertial masses of the pulsar and the companion star, m_1 and m_2 . A given theory $T(\beta', \beta'')$ can be rejected if the observationally derived curves do not intersect (within the experimental accuracies) at a single point. For PSR1913+16, values of β' and β'' lying above curve *a* in Fig. 3 prevent the curves for $\dot{\omega}$ and γ from intersecting anywhere. All tensor-biscalar theories with parameters in the region above line *a* can therefore be ruled out immediately, with high confidence⁶.

We further constrain the parametrized family of tensor-biscalar theories by fitting our data to the theory-dependent timing model introduced by Damour and Taylor⁶. In this approach one adjusts the (*a priori* unknown) masses m_1 and m_2 , and the theory parameters β' and β'' , instead of adjusting the independent phenomenological parameters $\dot{\omega}, \gamma, r, s, \dots$. The basic idea is a generalization of some solutions worked out^{12,21} within the framework of general relativity. The theory-dependent model is still firmly based on the work of Damour and Deruelle^{17,18}, but the post-keplerian quantities, which were free parameters in the phenomenological model, are now explicitly replaced in the fitting process by their predicted values in the theory $T(\beta', \beta'')$, expressed in terms of the keplerian parameters and the masses. In principle, one can fit this model directly to the timing data for a binary pulsar, obtaining simultaneous estimates of five keplerian parameters, the masses m_1 and m_2 , and the two theory parameters β' and β'' .

With the data sets available at present for PSR1913+16 and PSR1534+12, it is impossible to determine β' and β'' separately and uniquely by this method. Nevertheless, the results may be used to rule out large regions of otherwise viable theory space, especially when constraints based on the two pulsars are combined. We obtain separate limits for each pulsar by studying the variation of its goodness-of-fit statistic, χ^2 , for a large number of solutions in which the theory parameters β' and β'' are held fixed on a grid of trial values while the masses and keplerian parameters are adjusted for best fit to the data⁶. The results are presented in Fig. 3 in the form of closed contours representing the allowed (90% confidence) regions for β' and β'' for PSR1913+16 (*b*), and for PSR1534+12 (*d*, 68% and 90% confidence). The contours for the two pulsars have significantly different shapes because of the different mix of phenomenological effects appearing in their timing data. Essentially, the strongest test offered by PSR1913+16 is based on the measured values of $\dot{\omega}, \gamma$ and $\dot{P}_b$, whereas the tightest limits from PSR1534+12 involve $\dot{\omega}, \gamma, r$ and s . The area enclosed by the contours labelled *d* can be expected to decrease considerably in the near future, when additional measurements permit γ (and r) to be determined more accurately for this pulsar.

For completeness, Fig. 3 also contains a limit to β' derived from a test of the strong equivalence principle outlined in ref. 7 (see also refs 6, 8). This test relies on high-precision timing measurements of 'nonrelativistic' binary pulsars like PSR1953+29 (ref. 22) and PSR1855+09 (ref. 23), which have wide, nearly circular orbits. One attempts to detect a gravitational analogue of the Stark effect, in which the external perturbing force is the gravitational field of the Galaxy. More precisely, one looks for evidence that the neutron star and white dwarf in the PSR1855+09 system (only the former is strongly self-gravitating) are accelerated differently in the galactic field. This test generalizes to a

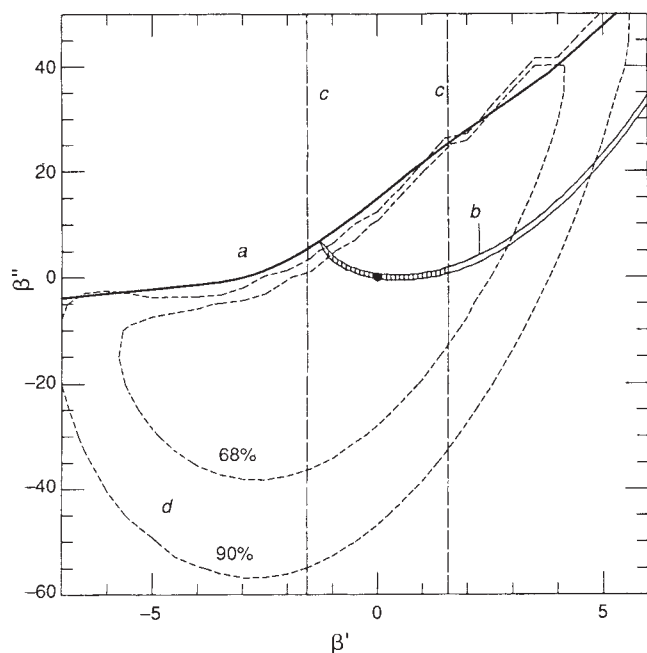


FIG. 3 Regions of the parametrized theory-space (β' , β'') lying above curve a are incompatible with the observed values of $\dot{\omega}$, γ and the keplerian parameters of the PSR1913+16 system. Contour b encloses the region allowed by our theory-dependent analysis of data for this pulsar (see text). Vertical lines (c) delimit the range of β' values consistent with the orbital parameters of PSR1855+09, on the basis of an upper limit for its violation of the strong equivalence principle⁷. Contours labelled d outline the region consistent with timing observations of PSR1534+12. All contours have been calculated for 90% confidence levels, except that an additional 68% contour is included for PSR1534+12 to help illustrate how the diagram may evolve when additional data are available. The intersection of regions allowed by all four tests is cross-hatched for emphasis, and the black dot at the origin corresponds to general relativity.

strong-field situation the weak-field, Solar-System test first proposed by Nordtvedt²⁴. The dashed vertical lines labelled c in Fig. 3 give the resulting limits on β' for PSR1855+09. (This limit is numerically weaker but more secure than a similar one based on PSR1953+29, because for PSR1855+09 all relevant parameters of the binary system have been directly measured from the timing data.)

Conclusions and prospects

Together, the results presented in Fig. 3 imply that if gravity is correctly described by a tensor-biscalar theory, the adjustable parameters of the theory are strongly constrained. Combining the limits provided by all three pulsars, we obtain the following constraints on the tensor-biscalar theory parameters (90% confidence level):

$$\begin{aligned} -1.1 < \beta' < 1.6 \\ -1 < \beta'' < 6 \end{aligned}$$

More precisely, β' and β'' must lie within the cross-hatched portion of contour b in Fig. 3. As the point (0, 0) lies well inside the acceptable region, we conclude that general relativity remains fully consistent with our measurements, and that Einstein's theory has passed several new, deep and sensitive experimental tests. The most important new fact is that among these tests, the ones associated with curves a , c and d all concern the quasi-stationary strong-field regime of relativistic gravity, without mixing of radiative effects.

To give a quantitative measure of the significance of our experimental limits on β' and β'' in terms of possible strong-field deviations from general relativity, we have calculated the range of variation of the effective parameters $\tilde{G}$, $\tilde{\gamma}$, $\tilde{\beta}$, entering the lagrangian for a two-neutron-star binary system with $m_1 = m_2 = 1.4M_\odot$ when β' and β'' run over the cross-hatched region in Fig. 3. We find

$$\begin{aligned} \tilde{G} &= G_{\text{Newton}}(1.00^{+0.14}_{-0.11}) \\ \tilde{\gamma} &= 1.00 \pm 0.25 \\ \tilde{\beta} &= 1.0^{+0.9}_{-0.7} \end{aligned}$$

Numerical simulations⁶ make us optimistic that these experimental results will be further improved. In particular, the simulations show that extended and/or improved observations of PSR1534+12 should in a few years reduce the size of contour d in Fig. 3, thereby yielding much tighter nonradiative strong-field tests. Moreover, there is good reason to expect that additional post-keplerian parameters will become measurable for these and other pulsars—including r and s separately for PSR1913+16, and $\dot{P}_b$ for PSR1534+12—and that these will lead to further theoretical constraints. The present upgrading of the Arecibo telescope promises to improve its sensitivity considerably, with consequences of great benefit to pulsar timing studies. Finally, we consider it likely that further pulsar searches now in progress will discover additional examples of binary pulsars in relativistically interesting orbits. $\square$

Received 27 September; accepted 3 December 1991.

- Will, C.M. *Theory and Experiment in Gravitational Physics* (Cambridge University Press, 1981).
- Shapiro, I.I. in *General Relativity and Gravitation, 1989* (eds Ashby, N., Bartlett, D.F. & Wyss, W.) 313–330 (Cambridge University Press, 1990).
- Taylor, J.H. *Proc. IEEE* **79**, 1054–1062 (1991).
- Hulse, R.A. & Taylor, J.H. *Astrophys. J.* **195**, L51–53 (1975).
- Wolszczan, A. *Nature* **350**, 688–690 (1991).
- Damour, T. & Taylor, J.H. *Phys. Rev. D* (in the press).
- Damour, T. & Schäfer, G. *Phys. Rev. Lett.* **66**, 2549–2552 (1991).
- Damour, T. & Esposito-Farèse, G. *Class. Quant. Grav.* (submitted).
- Taylor, J.H., Fowler, L.A. & McCulloch, P.M. *Nature* **277**, 437–440 (1979).
- Taylor, J.H. & Weisberg, J.M. *Astrophys. J.* **253**, 908–920 (1982).
- Weisberg, J.M. & Taylor, J.H. *Phys. Rev. Lett.* **52**, 1348–1350 (1984).

- Taylor, J.H. & Weisberg, J.M. *Astrophys. J.* **345**, 434–450 (1989).
- Damour, T. & Taylor, J.H. *Astrophys. J.* **366**, 501–511 (1991).
- Blandford, R. & Teukolsky, S.A. *Astrophys. J.* **205**, 580–591 (1976).
- Epstein, R. *Astrophys. J.* **216**, 92–100 (1977); erratum *Astrophys. J.* **231**, 644 (1979).
- Haugan, M.P. *Astrophys. J.* **296**, 1–12 (1985).
- Damour, T. & Deruelle, N. *Ann. Inst. H. Poincaré (Phys. théor.)* **44**, 263–292 (1986).
- Damour, T. & Deruelle, N. *Ann. Inst. H. Poincaré (Phys. théor.)* **43**, 107–132 (1985).
- Shapiro, I.I. *Phys. Rev. Lett.* **13**, 789–791 (1964).
- Will, C.M. & Zaglauer, H.W. *Astrophys. J.* **346**, 366–377 (1989).
- Taylor, J.H. in *General Relativity and Gravitation* (ed. MacCallum, M.A.H.) 209–222 (Cambridge University Press, 1987).
- Rawley, L.A., Taylor, J.H. & Davis, M.M. *Nature* **319**, 383–384 (1986).
- Ryba, M.F. & Taylor, J.H. *Astrophys. J.* **371**, 739–748 (1991).
- Nordtvedt, K. *Phys. Rev.* **170**, 1186–1187 (1968).

Atomic structure of single-stranded DNA bacteriophage Φ X174 and its functional implications

Robert McKenna*, Di Xia*, Peter Willingmann*[†], Leodevico L. Ilag*[‡], S. Krishnaswamy*[‡], Michael G. Rossmann*[§], Norman H. Olson*, Timothy S. Baker* & Nino L. Incardona[‡]

* Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, USA

[‡] Department of Microbiology and Immunology, Center for the Health Sciences, University of Tennessee, 858 Madison Avenue, Memphis, Tennessee 38163, USA

The mechanism of DNA ejection, viral assembly and evolution are related to the structure of bacteriophage Φ X174. The F protein forms a $T=1$ capsid whose major folding motif is the eight-stranded antiparallel β barrel found in many other icosahedral viruses. Groups of 5 G proteins form 12 dominating spikes that enclose a hydrophilic channel containing some diffuse electron density. Each G protein is a tight β barrel with its strands running radially outwards and with a topology similar to that of the F protein. The 12 'pilot' H proteins per virion may be partially located in the putative ion channel. The small, basic J protein is associated with the DNA and is situated in an interior cleft of the F protein. Tentatively, there are three regions of partially ordered DNA structure, accounting for about 12% of the total genome.

BACTERIOPHAGE Φ X174 is a small icosahedral virus that contains a single-stranded, closed circular DNA molecule with 5,386 nucleotide bases (for a recent review, see ref. 1). Of the 11 gene products, four (J, F, G and H) participate in the structure of the virion. There are 60 copies each of the J, F and G proteins and 12 copies of the H protein per virion²⁻⁴ (Table 1). Electron microscopy of shadowed or negatively stained particles⁵⁻⁸, as well as a three-dimensional reconstruction of frozen-hydrated virions (Fig. 1a), clearly show large spikes at the vertices of the icosahedral particles (N.H.O., T.S.B., P.W. and N.L.I., manuscript in preparation). These spikes are about 32 Å long and have a mean diameter of 70 Å. Edgell *et al.*³ have shown that these spikes, which can be removed from the capsid with 4 M urea treatment, are composed of G and H proteins. Assuming all 12 spikes are identical, then each spike will contain five G and one H protein². The capsid itself has an external diameter of roughly 260 Å, with a protein shell that is about 30 Å thick, and contains the F and J proteins as well as the single-stranded (ss) DNA. The phage recognizes a lipopolysaccharide (LPS) receptor on the outer cell membrane of *Escherichia coli*⁸⁻¹¹. Host range mutants, which determine which strains of *E. coli* are infected by Φ X174, map into the F, G and H genes¹²⁻¹⁵.

After attachment, which can occur at or above 15 °C, the

phage DNA and at least one of the H proteins are injected into the *E. coli* periplasmic space on raising the temperature to 37 °C. As this temperature shift produces an irreversible loss of infectivity, it is called the eclipse reaction. The transfer of phage DNA into the cell requires energy and may be coupled to synthesis of the complementary DNA strand¹⁶. Eclipse rate, as measured by loss of infectivity, can be modulated by deletion of certain bases in the viral genome¹⁷ as well as by mutations that map in genes F (L.L.I., J.K. Tuech, L. A. Beisner, R. A. Sumrada and N.L.I., manuscript in preparation) and H (ref. 18, and L.L.I. and N.L.I., manuscript in preparation). The lipid portion of the LPS receptor is required for the DNA injection process¹⁰. However, eclipse can also be initiated by incubation at 37 °C in the presence of 100 mM CaCl₂ (ref. 19). The eclipsed particles, which have a sedimentation coefficient of 90S, are sensitive to nucleases¹³. Electron microscopy^{20,21} shows DNA is an extrusion from one of the spikes when the LPS is used and from many spikes when Ca²⁺ is used to initiate eclipse.

In later stages of the viral life cycle, five molecules of F protein form 9S particles and five molecules of G protein form 6S particles. These pentamers together with the H protein and the B and D scaffolding proteins form 108S proheads. The ssDNA is then packaged into proheads along with 60 J proteins to form, after removal of the B proteins, a 132S complex^{22,23}. Conversion to the infective 114S particles involves removal of the D protein, requires divalent cations and may occur in the final stages of morphogenesis and cell lysis²⁴. Various temperature-sensitive (ts) and chain-termination mutants that block the assembly pathway have been isolated and characterized¹ (L.L.I. and N.L.I., manuscript in preparation). The small J protein is required for DNA packaging²³. They contain six lysines and six arginines, typical of other viral capsid proteins that have a role in nucleic acid packaging (see, for instance, ref. 25).

The phage Φ X174 was the first example of a biological system in which the complete nucleotide sequence of a DNA genome was successfully determined^{26,27} and shown to contain 11 genes with overlapping reading frames. Several similar phages have been isolated²⁸ of which only Φ X174, G4 (ref. 24) and S13 (ref. 30) have been sequenced. Comparisons of the protein sequences demonstrate that the spike protein G is the most variable. The H protein contains a much larger percentage of glycines and smaller percentage of aromatic residues than is usual for globular proteins. Such sequences often give rise to disordered peptide conformations and, hence, lack distinct structure³¹.

The infectious virion has a relative molecular mass of 6.2×10^6 (M_r 6,200K) (26% is DNA)³². Lysates of infected *E. coli* contain two components which can be separated by sedimentation through a sucrose gradient^{32,33}. The fast band contains the infectious 114S particles whereas the slow band contains noninfectious, 70S particles^{32,34,35}. The latter contains about 20% of the normal DNA content with all of the genome represented in

[†] Present addresses: Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK (P.W.); Bioinformatics Center, Department of Biotechnology, Madurai Kamaraj University, Madurai 625 021, India (S.K.).

[§] To whom correspondence should be addressed.

TABLE 1 Proteins in the virion

Protein	M_r	Number of amino acids	Amino-acid characteristics	Function
F	48,400	426	Normal	DNA injection and protection
G	19,050	175	Less conserved than other capsid proteins among related phages	Attachment
H	34,400	328	High Gly and low aromatic content	Attachment, DNA ejection, penetration ('the pilot protein')
J	4,200	37	Highly basic amino terminus	DNA packaging

the population of particles. Both types of particles can be crystallized into isomorphous monoclinic crystals with space group $P2_1$ and cell dimensions $a = 305.6$, $b = 360.8$, $c = 299.5$ Å, $\beta = 92.89^\circ$ that diffract at least to 2.6-Å resolution³³. The asymmetric unit of the crystal cell contains one complete particle. We report here the three-dimensional structure determination of Φ X174 and its implication for viral functions.

Protein structure

The structure determination of Φ X174 is summarized in Table 2 and will be described in detail elsewhere (R.M., D.X., P.W., L.L.I. and M.G.R., manuscript in preparation). The structure was initially interpreted at 3.5 Å resolution in terms of the amino-acid sequences of the F, J and G proteins³⁶, but the mutant Φ X174 used here has an arginine at position 216 in the F protein (Fig. 1*b*). An atomic model was built in an Evans and Sutherland PS390 graphics system with respect to a final 3.4 Å resolution map with the use of the program FRODO³⁷. The overall appearance of the atomic structure of Φ X174 is remarkably similar to that found by electron microscopy of unstained, frozen-hydrated virions (Fig. 1). The maximum outer radius is 165 Å measured along the 5-fold axes. Each of the 12 spikes consists of 5 G proteins which make only minimal contact with

the capsid itself (Fig. 2*d*, Table 3). The spikes have a pentagonal cross-section (Fig. 2*g*), with an average diameter of 70 Å, and a somewhat blunt extremity that extends about 32 Å above the mean capsid surface. The capsid has a number of protrusions on its surface, one of which extends to a radius of 138 Å at the 3-fold icosahedral axes, but at the 2-fold axes there is a depression in the surface that only extends to a radius of 111 Å. The protein coat has an average thickness of about 30 Å.

The capsid itself consists primarily of the F protein, with $T = 1$ (ref. 38) icosahedral symmetry, and shows protrusions on the 3-fold axes, in the centre of the triangle formed by adjacent 5- and 3-fold axes, and midway between the 5- and 3-fold axes. The F polypeptide was easily traced with confidence except for the first and last two residues which are disordered. The central motif of the F protein is an eight-stranded antiparallel β barrel similar to that found in many other virus coat proteins (Fig. 2*b*), although the usual ' α A' and ' α B' helices are missing (see, for example, ref. 39 for definition of secondary structure nomenclature). The organization of the β barrel with respect to the icosahedral symmetry is similar to that in other viruses: the β strands run roughly tangential to the viral surface and the B-C, D-E, F-G and H-I turns point toward the 5-fold axes. This β barrel projects into the nucleic-acid region of the virus. The contact between neighbouring β barrels is limited, occurring at the 5-fold vertices (Table 3), and, therefore, cannot be considered to be the dominating interactions required for assembly. Superposition of the F subunit β barrel onto VP1 of human rhinovirus 14 (HRV14)⁴⁰ and canine parvovirus (CPV)⁴¹ shows that 101 and 92 amino acids of the F protein can be topologically superimposed onto the corresponding β barrels of HRV14 and CPV with an r.m.s. deviation of 3.5 and 3.4 Å, respectively, for equivalenced C α coordinates. The insertions between the β strands (Fig. 2*a*) are almost as extensive as in CPV, accounting for more than 60% of the total F protein. These insertions form the exterior of the viral capsid and most of the contacts between neighbouring F subunits (Table 3). The insertions contain five α helices of which one has five-and-a-half turns. The long helix runs antiparallel to two short and nearly colinear helices from a 3-fold and a 2-fold related F subunit, respectively. The insertions also account for the protrusions on the viral surface.

The J protein is internal²³ and, because of its high content of basic residues, was expected to be associated with DNA.

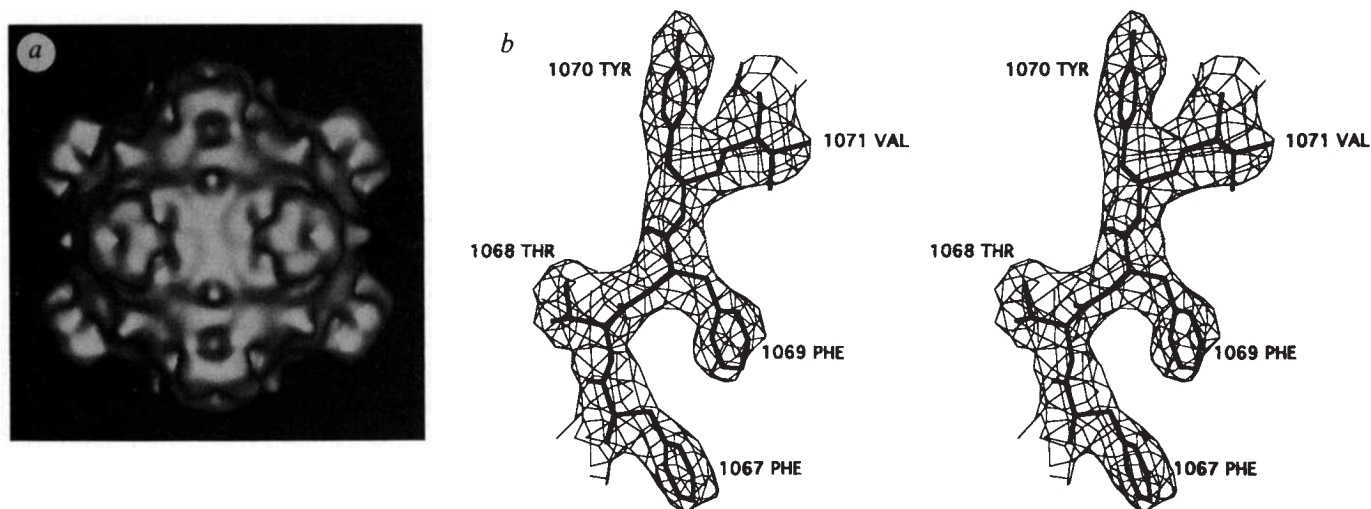


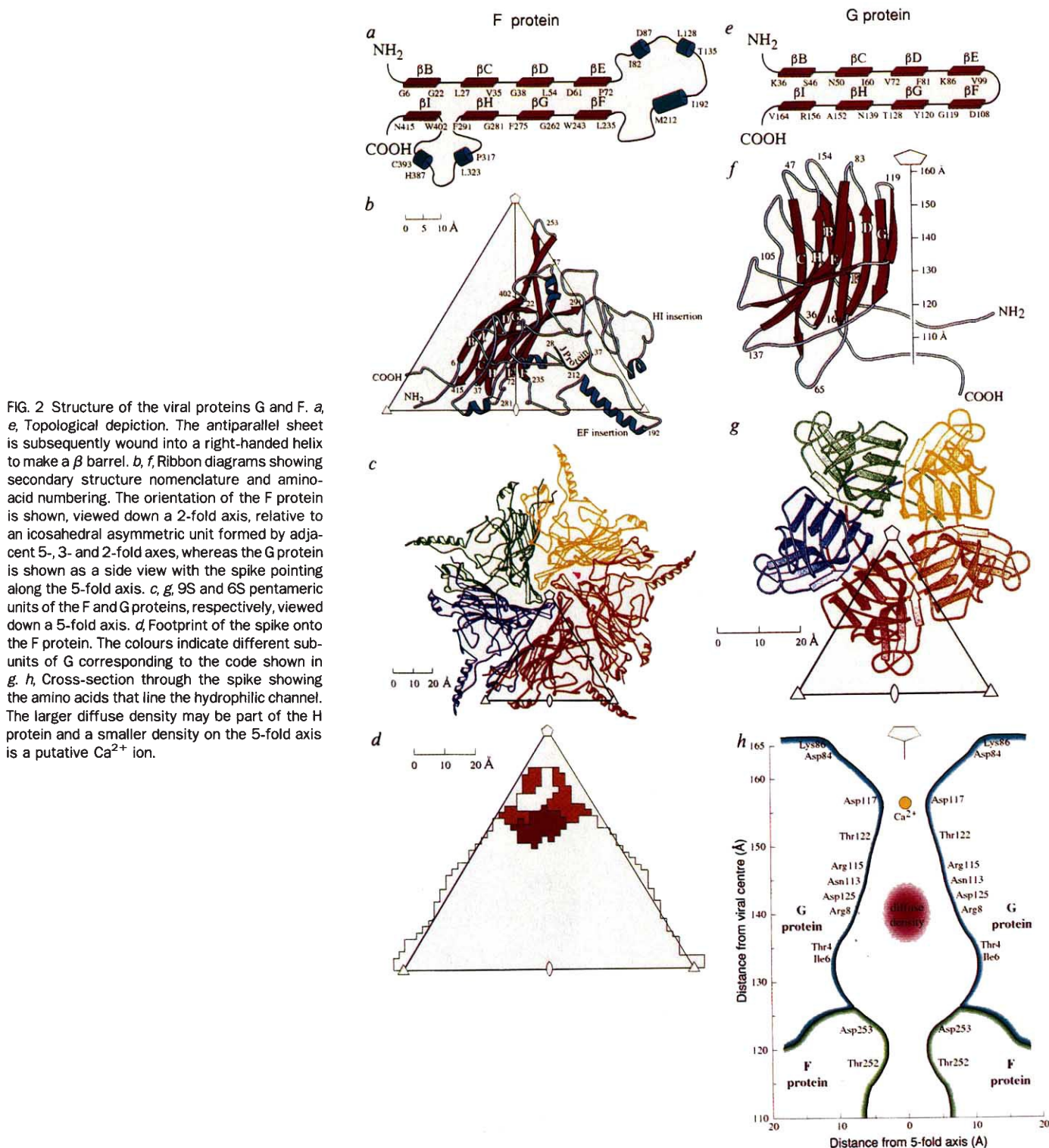
FIG. 1 *a*, Shaded surface representation of three-dimensional reconstruction of Φ X174 calculated from particle images recorded in the electron microscope. Unstained virus samples were frozen in a layer of vitreous ice and images were recorded under minimal electron dose conditions (~ 20 e/Å²). Twenty-five Φ X174 particle images were selected, their centres of density (origins) and orientation positions in the layer of ice were determined, and

reconstructions were calculated to 21 Å (see N.H.O., T.S.B., P.W. and N.L.I., manuscript in preparation, for details on procedures). The view is down an icosahedral 2-fold axis. *b*, Stereo view of fit of residues 67-FTFYV-71 in the β E strand of the F protein electron density⁶⁶. The amino-acid numbers in the F protein have been increased by 1,000.

However, carboxy-terminal residues 28-RLWYVGGQF-37 contain only one basic residue and are ordered. Their characteristic sequence was easy to recognize in the electron density. They make a loop on an interior cleft of the capsid protein and contact the E-F insertion. Many plant viruses have very basic, rather variable, amino-terminal extensions to their capsid proteins. These basic regions are almost invariably 'disordered' (for example, tomato bushy stunt virus⁴², southern bean mosaic virus⁴³ and satellite tobacco necrosis virus⁴⁴), lacking a unique

conformation. When there is no especially basic protein component in the capsid protein, there usually is a significant amount of polyamines to neutralize the nucleic acid⁴⁵. The disordered basic part of the J protein, therefore, is probably involved in DNA neutralization and packaging and may account for the difference of the Raman spectrum between ΦX174-DNA when in solution and in the virion⁴⁶.

The G polypeptide was traced over its entire length with complete confidence. Its structure is a somewhat complex β



barrel with its strands running roughly radially (Fig. 2*f, g*). It has considerable resemblance to the β -barrel domain of the F protein (Fig. 2*a, e*) and, indeed, to many other viral capsid proteins³⁹. The β turns, BC, DE, FG and HI, that point toward the 5-fold axes in many viral capsid proteins, are at the blunt extremity of the G protein spike. Of the 175 amino acids in the G protein, 73 residues can be topologically superimposed onto the F protein with a 3.0 Å r.m.s. difference in the α positions. There is a rather hydrophilic channel down the 5-fold axis of each spike (Fig. 2*h*), with a diameter which varies between 6 and 26 Å. There is diffuse electron density in this channel at a radius of about 135 Å. There is also another uninterpreted, but smaller, piece of electron density at a radius of 159 Å. The total cavity in the channel is large enough to accommodate at least

15 amino acids. The quality of the G-protein electron density is particularly surprising as the crystallographic environment of each spike is different and might, therefore, have asymmetrical lattice forces causing local deviation from icosahedral symmetry and, hence, blurring of electron density. Furthermore, the quality of the map is even more surprising because the spikes make only a few contacts with the capsid itself (Table 3).

There is no compelling evidence in the electron density map for the presence of the H protein. This is not surprising, given the unusually large number of glycines and the low percentage of aromatic residues in the H protein, which suggest that this protein may not have a distinct fold³¹. The absence of any defined structure for the H protein may be due to the effect of electron density averaging, to different orientations of the pro-

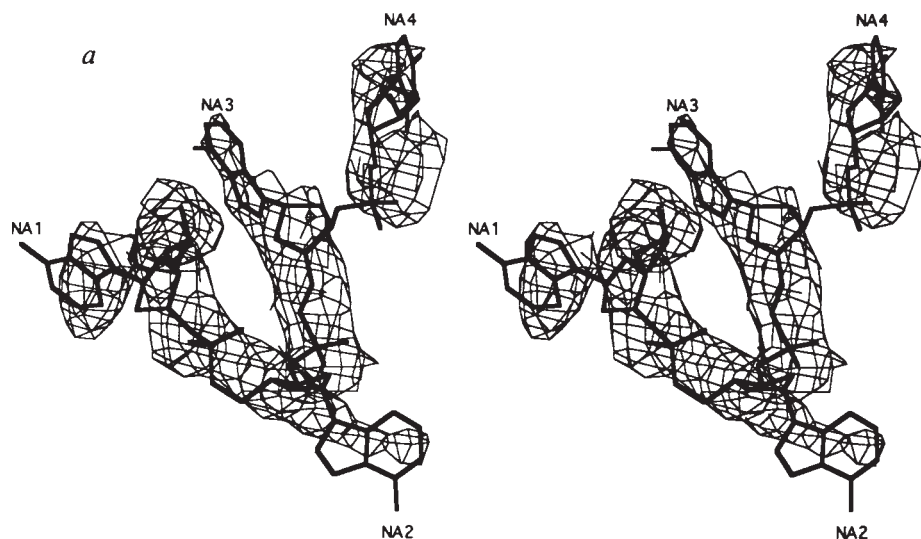


FIG. 3 *a*, Stereo view of fit of DNA bases NA1 to NA4 to the electron density. The choice of nucleotide bases shown is arbitrary. *b*, Stereo diagram of environment of the J protein (green). The DNA structure is in red and the F protein is in yellow. Orientation is given by the purple 5- and 2-fold symmetry axes. The polar residues on F and J have been identified. The amino-acid numbers have been increased by 1,000 and 3,000 for the F and J proteins, respectively. The nucleotides are numbered NA1 to NA11 although they occur in three noncontiguous sequences.

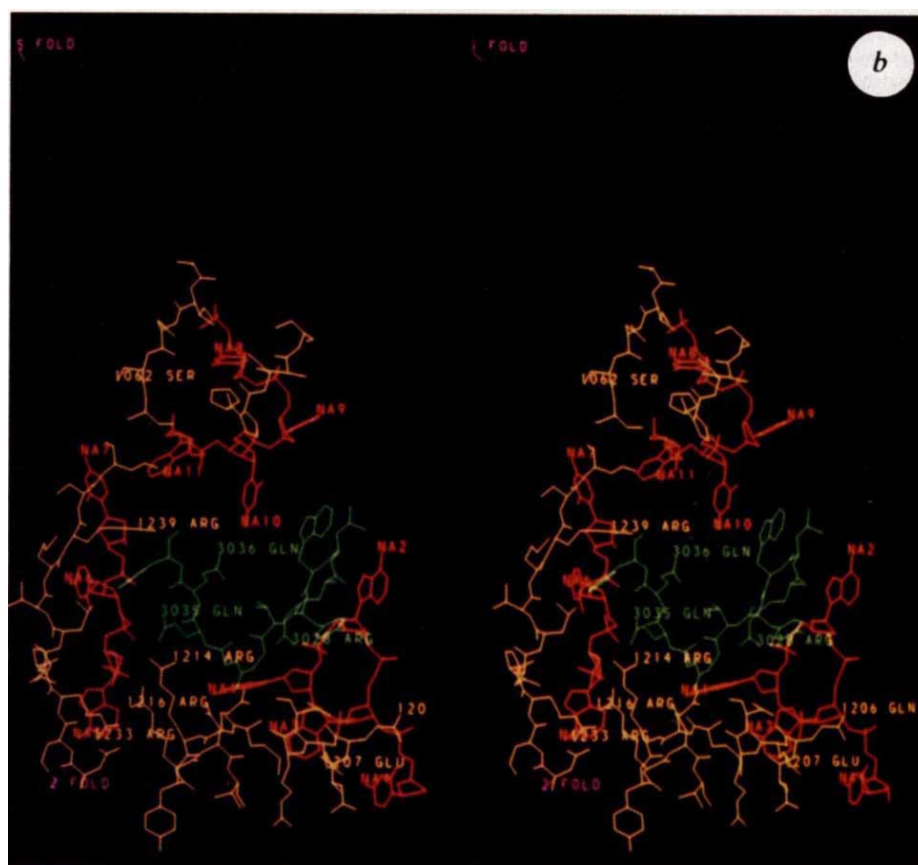


TABLE 2 X-ray diffraction data collection

Resolution range (Å)	Combined data (from 70S and 114S particles)		Data from 70S particles		Data from 114S particles	
	Number	%	Number	%	Number	%
∞–30.0	940	73	924	72	95	7
30.0–15.0	8,133	91	7,976	89	1,044	12
15.0–10.0	22,240	92	21,794	90	3,039	12
10.0–7.5	40,725	86	39,467	83	5,870	12
7.5–5.0	148,401	76	145,712	75	24,130	12
5.0–4.0	193,012	73	189,722	72	31,897	12
4.0–3.5	177,528	67	168,282	64	28,447	11
3.5–3.2	148,155	60	142,153	58	22,990	9
3.2–3.0	115,489	51	112,079	49	17,295	8
3.0–2.9	53,573	39	50,945	37	4,777	3
2.9–2.7	92,108	27	86,295	25	8,754	3
Independent reflections	1,000,304		965,349		148,338	
Number of films	299		279		20	
R_{merge}	12.90		12.04		12.49	

Number of observations ($F^2 > 2\sigma(F^2)$) for each data set and percentage of possible total.

R_{merge} is defined as

$$\frac{\sum_h \sum_i |(\langle I_h \rangle - I_{hi})|}{\sum_h \sum_i \langle I_h \rangle} \times 100$$

where $\langle I_h \rangle$ is the mean intensity of the i observations I_{hi} for reflection h . The virus was propagated, purified and crystallized as described by Willingmann *et al.*³³. Crystals of the partially empty, slower sedimenting particles were easiest to grow. A complete X-ray diffraction data set was collected for these crystals, whereas only a fraction of the data was collected from crystals of the faster-sedimenting virus. In addition, partial data sets were collected for two heavy atom derivatives of the slower-sedimenting particles. X-ray diffraction data were obtained from non-oriented crystals as described by Willingmann *et al.*³³. Data from crystals of full 114S particles and from partially empty 70S particles were combined for the initial structure determination. Later, these were separated to determine the presence of ordered DNA in the crystal structure. A self-rotation function⁶¹, calculated initially at 10 Å resolution, showed icosahedral symmetry and gave an orientation of a particle in the asymmetric unit. This was refined using a locked rotation function⁶² at 3.5 Å resolution. Packing considerations indicated that a particle position must be close to $(\frac{1}{2}, y, \frac{1}{2})$ in the unit cell. Several initial phasing sets, for data between 50 and 25 Å resolution, derived from electron microscopy image reconstructions and spherical shell models, failed to give useful phase extensions. In hindsight, these failures were probably caused by a lack of knowledge of the precise particle position. An insightful observation (J. E. Johnson, personal communication) was made that the overall external shape, appearance and size of ΦX174 was rather similar to that of cowpea mosaic virus (CpMV). As the atomic structure of CpMV was known⁶³, it could be used as a search model at 12 Å resolution. This yielded a particle position at (0.2440, 0.2500, 0.2480) which was 2 Å away from any previously used centre. By contrast, the structure factors derived from the ΦX174 electron microscopy reconstruction had essentially zero amplitude beyond 25 Å resolution. Phases were then slowly extended by averaging the 60 noncrystallographically related electron densities in the viral envelope (see, for instance, ref. 64). The particle centre, redetermined several times during the phase extension process, refined to a final position of (0.2505, 0.2500, 0.2505). The markedly nonspherical particle envelope was also redetermined at regular intervals⁶⁵. Some heavy atom derivative data had been collected (R.M., D.X., P.W., L.L.I. and M.G.R., manuscript in preparation). Difference Fourier at various resolutions showed that the Babinet opposite solution had been obtained^{41,56}. The phases were, therefore, flipped ($\alpha \rightarrow \pi + \alpha$) before plotting the final electron density map. The CpMV phasing model had not been used in the usual molecular replacement manner to solve the ΦX174 structures as the atomic structures of CpMV and ΦX174 are almost entirely different. The CpMV model served only to obtain a set of initial, very low resolution, structure factors. These structure amplitudes were used in an R -factor search for the particle centre that had phases modulated by the envelope shape but not its contents.

tein at each spike or, most likely, to different folds of each polypeptide. Nevertheless, denaturing gels of virus dissolved from crystals show clear evidence of its presence³³. In light of the 'pilot' function of the H protein (Table 1), it is possible that the H protein would bind to the outside extremity of the spikes. In the crystal, the spikes interdigitate, which leaves little room for a 35K protein at the vertices of the spikes. It is unlikely that DNA accounts for the diffuse density in the large cavity on the 5-fold axis of each spike as there is insufficient space for a strand of DNA to enter and exit this cavity. Genetic evidence suggests that Asn 113 of the G protein, which is in the channel facing the diffuse density (Fig. 2h), interacts with Ala 68 of the H protein (L.L.I., *et al.*, manuscript in preparation). Therefore, the diffuse density might be part of the H protein. However, it is not clear whether the remainder of the H protein is internal to the capsid or external to the virus, though both possibilities would be consistent with the electron microscope observation that DNA is ejected through the spikes²¹ and that the H protein is injected into *E. coli* with DNA⁴⁷.

DNA structure and G protein ion channel

A difference electron density map was calculated between the partial diffraction data (Table 2) of the full 114S particles, F_{114S} , and the mostly empty 70S particles, F_{70S} , based on the phases, α_{combined} , determined for the 3.4 Å resolution map. This shows regions where there is ordered structure not present in the electron density map of the 70S particles. This structure is

probably attributable to DNA (Fig. 3a) but it could be also a part of the H protein. The basic protein environment and characteristic electron density strongly suggested this to be nucleic acid. As in bean pod mottle virus⁴⁸ and CPV⁴¹, some of the nucleic acid is bound to inside cavities of the icosahedral shell and, therefore, has itself attained some icosahedral order. The most extensive ordered DNA structure occurs in three regions of the cavities between the F protein β barrel and the rest of the capsid, including a region close to the ordered part of the J protein. A total of 11 nucleotides per icosahedral asymmetric unit could be built into the electron density while maintaining standard dihedral angles⁴⁹. There are extensive polar interactions of the DNA backbone with the F and J proteins including Arg 216, Arg 233 and Arg 420 in the F protein as well as Arg 28 in the J protein (Fig. 3b). The order of the protein in this region is greater in the 114S than in the 70S particles. An electron density map using $(F_{114S} - kF_{70S})e^{i\alpha_{\text{combined}}}$, where $k = 0.9$, showed the DNA density to have the same height as the protein. Full occupancy of DNA would be expected for $k = 0.5$ (ref. 50), implying that the ordered DNA is present in about one out of five icosahedral positions.

A large negative peak occurs in the difference map in the G protein channel at the site of the smaller piece of diffuse density, suggesting that a cation is bound here in the 70S particles. This density is liganded by the five symmetry-related Asp117 residues. Hence, binding of a putative Ca^{2+} ion at this site may trigger DNA ejection^{19–21} through the ion channel. Additional density

TABLE 3 Number of amino acid contacts of reference subunit with other protein subunits or DNA

	Reference subunits				
	F _{barrel} [*]	F _{else} [†]	G	J	DNA
F _{barrel}	7 (F-G turn)	13	7	1	8
F _{else}	11(βD, βG)	78 (44 to E-F insertion) (24 H-I insertion)	5	8	6
G	4 (F-G turn) (D-E turn)	4 (H-I insertion)	56	0	0
J	1	17 (E-F insertion)	0	0	2
DNA	21 (8 βF)	13	0	2	—

Contacts are defined as any amino acid which approaches another amino acid to within 3.8 Å.
^{*}F_{barrel} is defined as all amino acids in the F protein β barrel excluding the E-F and H-I insertions.
[†]F_{else} is defined as all other amino acids excluded from F_{barrel}.

in the full 114S particles occurs also on the inside of the F capsid but around the 5-fold axis, which could be either DNA or more ordered H protein poised for ejection through the channel in the spike. In addition, the J protein shows a higher degree of order in the 114S particles.

Mutations controlling infectivity stages

Both ts and cold-sensitive (cs) mutants have been selected and characterized. The cs mutations at residues Arg 216 and Arg 233 in the F protein affect the eclipse rate and are also implicated in functional interactions with DNA¹⁷ (L.L.I., *et al.*, manuscript in preparation). One of the more highly ordered segments of DNA structure is sandwiched between these residues and the ordered region of the J packaging protein (Fig. 3). Because a change of the above arginines to cysteines or histidine should alter protein-DNA interactions, it is reasonable to expect that these and other residues adjacent to this segment of DNA should have a role in determining the rate at which infectivity is lost because of DNA ejection.

Nearly all ts mutants have defects in late stages of viral infection, presumably in the assembly process (L.L.I. and N.L.I., manuscript in preparation). The mutations in the G and F proteins map primarily to surface features on the virus (Table 4). In addition, there is one mutation on the inside cavity of the spike facing the 5-fold axes. These mutations are mostly of the type that tend to destabilize the local protein environment⁵¹, thus explaining why elevated temperatures might disrupt assembly.

FIG. 4 Alignment of amino-acid sequences based on structure superpositions (single-letter amino-acid code). Shown are only those residues in the β-barrel structures. Residues that are boxed are conserved in at least two sequences. The F and G proteins are aligned with each other as well as with the similar virus β barrels found in canine parvovirus (CPV), VP1 of HRV14, the small protein of CpMV and the protruding domain of TBSV. Note that the helical hand of all these structures is the same (see Fig. 2a and e) contrary to the claim by Gibson and Argos⁶⁷ concerning the TBSV protruding domain. The orientation and position of the G protein in ΦX174 are similar to that of the small protein in CpMV accounting for the similar external appearance of these two viruses (Fig. 1).

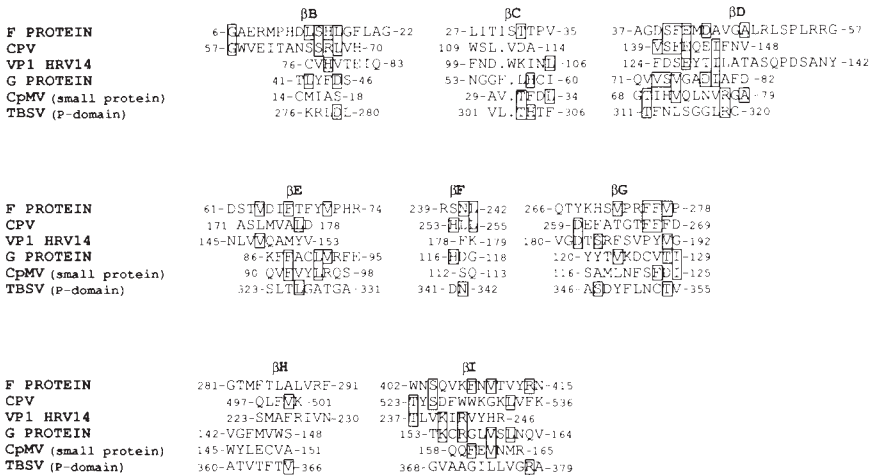


TABLE 4 Sites of mutations

Location of ts mutations that may block folding or assembly*	
G Protein	
(1) On spike surface	S39 → L(βB), G52 → S(βC), P100 → L(βE), P100 → S(βE), S148 → F(βH)
(2) On spike internal cavity	T4 → A (amino end), H116 → Y(βF)
F Protein	
(1) On exterior surface of the capsid	A91 → V (E-F loop), P120 → S (E-F loop), P148 → H (E-F loop), P175 → L (E-F loop), G337 → S (H-I loop), L384 → I (H-I loop)
(2) On interior surface of the capsid	H16 → Y(βB), M43 → I(βD)
Location of cs mutations that may block eclipse/DNA injection	
F Protein	R216 → H (near J), R216 → C (near J), R233 → C (near J) in interior cleft in close proximity to J

* L.L.I. and N.L.I., manuscript in preparation.

Evolution

Comparisons of the protein sequences of ΦX174 and G4 show the greatest variability in the extremity of the G protein and the greatest conservation in the contacts around the 5-fold axes in the F protein (particularly loop F-G). The higher rate of accepted mutation in the G protein would suggest that its function (perhaps formation of the channel through which the DNA is ejected) is a relatively new feature, whereas the capsid may have evolved earlier to a more stable structure. Superposition of tertiary protein structures permits amino-acid alignments when there is no readily recognizable memory of a common primordial amino-acid sequence⁵². The alignment of the virus β-barrel fold in the F and G proteins of ΦX174 with some other viral capsid proteins, as well as the protruding domain of tomato bushy stunt virus, is shown in Fig. 4. There is a tendency to conserve hydrophobic residues in the β barrel, but there is no obvious specific sequence, as sometimes occurs, by which the virus capsid fold might be readily recognized. The degree of similarity of the ΦX174 protein folds to the other capsid protein structures is similar to that between, for instance, HRV14 VP1 and VP2 which occur in the same virus capsid.

The considerable similarity of the F protein to other viral capsid proteins suggests that many viral capsid proteins have probably diverged from a common ancestral structure⁵³. The F gene has, over time, captured insertions which are important for forming the capsid exterior. As the virion exterior determines the interaction with its host, and as the host presumably already contains molecules that modulate its metabolism, such gene

fusion processes in viral evolution may be important in the function of the resultant viral particle.

The eight-stranded antiparallel virus fold has now been found in many RNA and DNA viral capsids³⁹, yet its presence is not entirely universal (for example in the plant RNA tobacco mosaic virus^{54,55}, in the RNA phage MS2⁵⁶ and in the animal RNA Sindbis virus⁵⁷). Hence, the eight-stranded antiparallel β -barrel

fold is not a prerequisite for assembly of icosahedral structures⁵⁸. Furthermore, the Φ X174 structure, in contrast to other icosahedral viruses, shows only little contact of the β -barrel fold with neighbouring subunits, as it essentially projects into the DNA core (Table 3). It has also been suggested^{59,60} that the fold might be conserved to modulate viral stability by virtue of its hydrophobic interior. $\square$

Received 5 August; accepted 31 October 1991.

1. Hayashi, M., Aoyama, A., Richardson, D. L. Jr & Hayashi, M. N. in *The Bacteriophages (The Viruses)* Vol. 2 (ed. Calendar, R.) 1-71 (Plenum, New York, 1988).
2. Burgess, A. B. *Proc. natn. Acad. Sci. U.S.A.* **64**, 613-617 (1969).
3. Edgell, M. H., Hutchison, C. A. III & Sinsheimer, R. L. *J. molec. Biol.* **42**, 547-557 (1969).
4. Siden, E. J. & Hayashi, M. *J. molec. Biol.* **89**, 1-16 (1974).
5. Hall, C. E., Maclean, E. C. & Tessman, I. *J. molec. Biol.* **1**, 192-194 (1959).
6. Thomas, W. J. & Horne, R. W. *Virology* **15**, 1-7 (1961).
7. Stouthamer, A. H., Daems, W. T. & Eigner, J. *Virology* **20**, 246-250 (1963).
8. Brown, D. T., Mackenzie, J. M. & Bayer, M. E. *J. Virol.* **7**, 836-846 (1971).
9. Incardona, N. L. & Selvidge, L. *J. Virol.* **11**, 775-782 (1973).
10. Jazwinski, S. M., Lindberg, A. A. & Kornberg, A. *Virology* **66**, 268-282 (1975).
11. Feige, U. & Stirm, S. *Biochem. biophys. Res. Commun.* **71**, 566-573 (1976).
12. Sinsheimer, R. L. *Prog. Nucleic Acid Res. molec. Biol.* **8**, 115-169 (1968).
13. Newbold, J. E. & Sinsheimer, R. L. *J. Virol.* **5**, 427-431 (1970).
14. Weisbeek, P. J., Van de Pol, J. H. & Van Arkel, G. A. *Virology* **52**, 408-416 (1973).
15. Dowell, C. E., Jansz, H. S. & Zandberg, J. *Virology* **114**, 252-255 (1981).
16. Newbold, J. E. & Sinsheimer, R. L. *J. molec. Biol.* **49**, 49-66 (1970).
17. Incardona, N. L. & Müller, U. R. *J. molec. Biol.* **181**, 479-486 (1985).
18. Doniger, J. & Tessman, I. *Virology* **39**, 389-394 (1969).
19. Incardona, N. L. *J. Virol.* **14**, 469-478 (1974).
20. Yazaki, K. *J. virol. Meth.* **2**, 159-167 (1981).
21. Mano, Y., Kawabe, T., Komano, T. & Yazaki, K. *Agric. Biol. Chem.* **46**, 2041-2049 (1982).
22. Fujisawa, H. & Hayashi, M. *J. Virol.* **23**, 439-442 (1977).
23. Aoyama, A., Hamatake, R. K. & Hayashi, M. *Proc. natn. Acad. Sci. U.S.A.* **78**, 7285-7289 (1981).
24. Mukai, R., Hamatake, R. K. & Hayashi, M. *Proc. natn. Acad. Sci. U.S.A.* **76**, 4877-4881 (1979).
25. Liljas, L. *Prog. Biophys. molec. Biol.* **48**, 1-36 (1986).
26. Sanger, F. *et al. Nature* **265**, 687-695 (1977).
27. Shaw, D. C. *et al. Nature* **272**, 510-515 (1978).
28. Godson, G. N. in *The Single-Stranded DNA Phages* (eds Denhardt, D. T., Dressler, D. & Ray, D. S.) 103-112 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1978).
29. Godson, G. N., Fiddes, J. C., Barrell, B. G. & Sanger, F. in *The Single-Stranded DNA Phages* (eds Denhardt, D. T., Dressler, D. & Ray, D. S.) 51-86 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1978).
30. Lau, P. C. K. & Spencer, J. H. *Gene* **40**, 273-284 (1985).
31. Huber, R. & Bennett, W. S. Jr *Biopolymers* **22**, 261-279 (1983).
32. Sinsheimer, R. L. *J. molec. Biol.* **1**, 37-42 (1959).
33. Willingmann, P. *et al. J. molec. Biol.* **212**, 345-350 (1990).
34. Eigner, J., Stouthamer, A. H., van der Sluys, I. & Cohen, J. A. *J. molec. Biol.* **6**, 61-84 (1963).
35. Weisbeek, P. J., Van de Pol, J. H. & Van Arkel, G. A. *Virology* **48**, 456-462 (1972).
36. Sanger, F. *et al. J. molec. Biol.* **125**, 225-246 (1978).
37. Jones, T. A. *J. appl. Crystallogr.* **11**, 268-272 (1978).
38. Caspar, D. L. D. & Klug, A. *Cold Spring Harb. Symp. quant. Biol.* **27**, 1-24 (1962).
39. Rossmann, M. G. & Johnson, J. E. *A. Rev. Biochem.* **58**, 533-573 (1989).
40. Rossmann, M. G. *et al. Nature* **317**, 145-153 (1985).

41. Tsao, J. *et al. Science* **251**, 1456-1464 (1991).
42. Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. *Nature* **276**, 368-373 (1978).
43. Abad-Zapatero, C. *et al. Nature* **286**, 33-39 (1980).
44. Liljas, L. *et al. J. molec. Biol.* **159**, 93-108 (1982).
45. Cohen, S. S. & McCormick, F. P. *Adv. Virus Res.* **24**, 331-387 (1979).
46. Benevides, J. M., Stow, P. L., Ilag, L. L., Incardona, N. L. & Thomas, G. J. Jr *Biochemistry* **30**, 4855-4863 (1991).
47. Jazwinski, S. M., Marco, R. & Kornberg, A. *Virology* **66**, 294-305 (1975).
48. Chen, Z. *et al. Science* **245**, 154-159 (1989).
49. Saenger, W. in *Principles of Nucleic Acid Structure* (ed. Cantor, C. R.) 51-104 (Springer, New York, 1984).
50. Chapman, M. S., Minor, I., Rossmann, M. G., Diana, G. D. & Andries, K. *J. molec. Biol.* **217**, 455-463 (1991).
51. Argos, P. *et al. Biochemistry* **18**, 5698-5703 (1979).
52. Rossmann, M. G., Moras, D. & Olsen, K. W. *Nature* **250**, 194-199 (1974).
53. Matthews, B. W. & Rossmann, M. G. *Meth. Enzym.* **115**, 397-420 (1985).
54. Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. *Nature* **276**, 362-368 (1978).
55. Namba, K. & Stubbs, G. *Science* **231**, 1401-1406 (1986).
56. Valegård, K., Liljas, L., Fridborg, K. & Unge, T. *Nature* **345**, 36-41 (1990).
57. Choi, H. K. *et al. Nature* **354**, 37-43 (1991).
58. Ladenstein, R. *et al. J. molec. Biol.* **203**, 1045-1070 (1988).
59. Smith, T. J. *et al. Science* **233**, 1286-1293 (1986).
60. Rossmann, M. G. *Proc. natn. Acad. Sci. U.S.A.* **85**, 4625-4627 (1988).
61. Rossmann, M. G. & Blow, D. M. *Acta crystallogr.* **15**, 24-31 (1962).
62. Tong, L. & Rossmann, M. G. *Acta crystallogr.* **A46**, 783-792 (1990).
63. Stauffer, C. V. *et al. in Crystallography in Molecular Biology* (eds Moras, D., Drenth, J., Strandberg, B., Suck, D. & Wilson, K.) 293-308 (Plenum, New York, 1987).
64. Rossmann, M. G. *Acta crystallogr.* **A46**, 73-82 (1990).
65. Rossmann, M. G. *et al. J. appl. Crystallogr.* (in the press).
66. Smith, T. J. *J. appl. Crystallogr.* **23**, 141-142 (1990).
67. Gibson, T. J. & Argos, P. *J. molec. Biol.* **212**, 7-9 (1990).

ACKNOWLEDGEMENTS. We thank T. Smith and M. Chapman for discussions, the many participants in data collection trips (including M. Agbandje, J. Bibler, M. Chapman, H.-K. Choi, V. Giranda, A. Hadfield, W. Keller, M. Oliveira, A. Prongay, T. Smith, L. Tong and H. Wu) and the support at the various synchrotron facilities (the National Synchrotron Light Source at Brookhaven, the Cornell High Energy Synchrotron Source, the Scientific Research Council synchrotron in Daresbury, the Deutsches Elektronen-Synchrotron in Hamburg and the Stanford Synchrotron Research Laboratory) and H. Prongay, S. Wilder and M. O'Neil for help in preparation of the manuscript. The work was supported by a NSF grant and a Lucille P. Markey Charitable Trust grant for the development of structural studies at Purdue (to M.G.R.), an NSF grant (to N.L.I.), a travel award from the Institute for Biological Recognition and Catalysis (to L.L.I.), a NIH grant (to T.S.B.) and a postdoctoral fellowship from the Deutsche Forschungsgemeinschaft (to P.W.). Unrefined coordinates have been deposited with the Brookhaven Protein Data Bank.

LETTERS TO NATURE

Spatial distribution of γ -ray bursts observed by BATSE

C. A. Meegan*, G. J. Fishman*, R. B. Wilson*,
W. S. Paciesas†, G. N. Pendleton†, J. M. Horack*,
M. N. Brock* & C. Kouveliotou‡§

* Space Science Laboratory, NASA/Marshall Space Flight Center, Huntsville, Alabama 35812, USA

† Department of Physics, University of Alabama in Huntsville, Huntsville, Alabama 35899, USA

‡ Universities Space Research Association

THE nature of the sources of cosmic γ -ray bursts is a long-standing problem in astrophysics. Lack of knowledge of their true spatial distribution and of their intrinsic brightness has hampered theoretical understanding of these enigmatic events. The Burst and Transient Source Experiment on the Compton Gamma-Ray Observatory has been detecting bursts at the rate of about one a day, and we report here an analysis of 153 events. The number versus intensity distribution does not follow the $-3/2$ power law expected for a spatially extended homogeneous distribution of

sources, but at the same time the angular distribution is isotropic within statistical limits. Taken together, these results are inconsistent with the spatial distribution of any known population of galactic objects, but may be consistent with the bursts being at cosmological distances.

The Burst and Transient Source Experiment (BATSE) is described by Fishman *et al.*¹ It includes eight NaI (TI) detectors, 1.27 cm thick by 50.8 cm diameter, sensitive to γ -rays in the energy range of 20 keV to 2 MeV. Directions to the burst sources are determined by comparing count rates from the separate detectors. The requirement that the relative detector count rates yield a consistent location provides an effective discriminator for false triggers. With a flux threshold of $\sim 10^{-7}$ erg cm⁻², BATSE combines unprecedented sensitivity to weak bursts with the capability of determining locations of all detected bursts in a single experiment. Bursts are triggered on-board when the count rate in the 50-300-keV energy range increases by 5.5 σ or more in at least two detectors simultaneously. The rates are tested on three timescales independently: 64 ms, 256 ms and 1,024 ms. In the period of 193 days since activation of the instrument on 21 April 1991 until 31 October 1991, a total of 153 bursts have been detected and located. This corresponds to a full sky rate of ~ 800 bursts per year, a factor of ~ 6 greater than seen by previous sensitive instruments².

§ On leave from University of Athens, Greece.

The intensity distribution of the bursts provides information on the radial distribution of the sources. Figure 1a shows a plot of the integral number of bursts brighter than a specified peak intensity for 140 analysed BATSE bursts; peak rates have not yet been calculated for thirteen bursts because of a data processing lag. The $C_{\max}$ intensity is expressed as the peak count rate divided by the trigger threshold count rate³ $C_{\min}$. The use of this parameter removes biases due to varying trigger thresholds and other systematic effects such as atmospheric scatter that are apparent in traditional $\log N(>S)$ - $\log S$ plots where S is the fluence (erg cm^{-2}) of the bursts. A homogeneous distribution of sources would follow the $-3/2$ power law shown in the figure. The data indicate a much flatter distribution, which is best characterized as a deficit in the number of weak bursts compared with what would be expected for a homogeneous distribution. An equivalent representation is the $V/V_{\max}$ distribution³, where $V/V_{\max} = (C_{\max}/C_{\min})^{-3/2}$. In this ratio, V is then the volume contained by a sphere extending to the location of the burst, and $V_{\max}$ is the volume of a sphere extending to the maximum distance at which the same burst would still be detectable by the instrument. The average value of $V/V_{\max}$ provides a statistical test of the hypothesis that the observed bursts are drawn from a spatially homogeneous sample. For a homogeneous distribution of sources, $V/V_{\max}$ is uniformly distributed between 0 and 1, with an average $\langle V/V_{\max} \rangle = 1/2$. The $V/V_{\max}$ distribution is presented in Fig. 1b and has $\langle V/V_{\max} \rangle = 0.348 \pm 0.024$. Several other recent experiments have obtained similar results for the intensity distribution of weak bursts^{4,5}. BATSE has not been operating long enough to sample the stronger but rarer bursts that have been found by earlier instruments to follow the

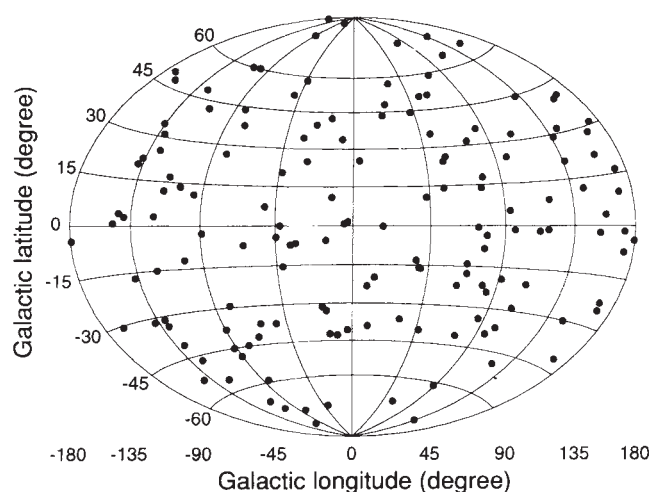


FIG. 2 The angular distribution of 153 bursts in galactic coordinates. There is no statistically significant deviation from isotropy.

$-3/2$ power law⁶⁻⁸. Although the $V/V_{\max}$ distribution is an excellent test for homogeneity, it obscures important information relevant to model fitting. Specifically, the strong peak in the lowest $V/V_{\max}$ bin corresponds to all bursts above $C_{\max}/C_{\min} = 4.6$. From Fig. 1a, it is clear that even these bursts are less numerous than would be expected from observations at higher intensities.

The angular distribution of 153 bursts is shown in Fig. 2. There is no apparent clustering of bursts along the galactic plane or toward the Galactic Centre. There is also no indication of excesses in the directions to nearby galaxies or clusters of galaxies, including the Virgo cluster. Hartmann and Epstein⁹ have promoted the use of the dipole and quadrupole moments to characterize angular distributions of bursts. A measure of the dipole moment in galactic coordinates is $\langle \cos \theta \rangle$, where θ is the angle between the burst source and the Galactic Centre. For BATSE bursts, $\langle \cos \theta \rangle = -0.002 \pm 0.006$, as compared with 0 ± 0.046 for an isotropic distribution. A measure of the quadrupole moment in galactic coordinates is $\langle \sin^2 b \rangle$, where b is the galactic latitude. For BATSE bursts, $\langle \sin^2 b \rangle = 0.310 \pm 0.006$, compared with 0.333 ± 0.023 for an isotropic distribution. The quoted errors for the measured values are the instrumental errors due to inaccuracy of the burst locations only. The quoted errors for the isotropic model distribution are the standard statistical errors for a sample of 153 events. Coordinate-free measurements of the moments are also statistically consistent with isotropy.

The BATSE results can be corrected for nonuniform sky exposure due to Earth blockage and time intervals when the trigger is disabled, such as during South Atlantic anomaly passages. The only anisotropy in the calculated exposure map is a 30% enhancement towards the celestial poles relative to the Equator. There is no appreciable dependence of exposure on galactic latitude or longitude. The corrected value of $\langle \cos \theta \rangle$ is 0.01; the corrected value of $\langle \sin^2 b \rangle$ is 0.31. Instrumental and statistical errors are the same as for the uncorrected case.

The accuracy of the burst locations is limited at present by systematic effects, primarily errors in the angular response of the detectors at high angle of incidence, and approximations in the atmospheric scattering corrections. Statistical errors contribute $\sim 10^\circ$ at the trigger threshold and average 3.5° for all bursts. Observations of 41 solar flares¹⁰ yield a mean error of 5.6° . Four bursts have been independently located using the interplanetary network^{11,12} and COMPTEL¹³. BATSE errors for these bursts are $\sim 6^\circ$. The location errors will be reduced as systematic effects are understood.

These results indicate that the burst sources are distributed isotropically but not homogeneously. No known galactic objects

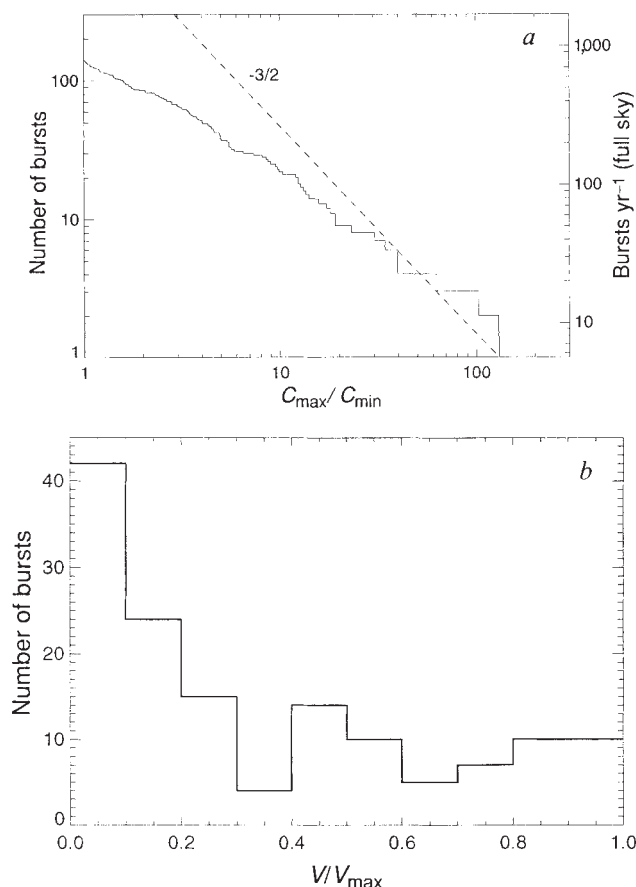


FIG. 1 a, Integral number distribution of 140 bursts as a function of peak rate. A $-3/2$ power law is expected for a homogeneous distribution of sources. The full sky rate is ~ 800 bursts per year. b, $V/V_{\max}$ distribution for 140 bursts. The average $V/V_{\max}$ is 0.348 ± 0.024 . A uniform distribution is expected for a homogeneous distribution of sources.

have such a distribution. Disk models produce unacceptable quadrupole moments for distant sources or unacceptable $\langle V/V_{\max} \rangle$ for nearby sources¹⁴. Galactic halo distributions¹⁵ must be at least 50 kpc in radius without strong central condensation to satisfy the BATSE limits on the dipole moment. Populating a halo with neutron stars (the currently favoured candidates) may prove difficult. Nearby extragalactic models will have difficulty with the lack of correlations with M31 and the Virgo cluster¹⁶. The isotropy and $V/V_{\max}$ results can be reproduced by cosmological models^{17,18}, which require isotropy but yield $V/V_{\max} < 0.5$ because of redshift effects. These models will require further development to address details of the burst phenomenon, such as energy spectra, occurrence rate and luminosity. $\square$

Received 22 November; accepted 2 December 1991.

1. Fishman, G. J. et al. in *Proc. Gamma Ray Observatory Science Workshop* (ed. Johnson, W. N.) 39–50 (NASA/GSFC, Greenbelt, 1989).
2. Golenetskii, S. *Adv. Space Res.* **8**(2), 653–657 (1988).
3. Schmidt, M., Higdon, J. C. & Hueter, G. *Astrophys. J.* **329**, L85–87 (1988).
4. Atteia, J.-L. et al. *Nature* **351**, 296–298 (1991).
5. Ogasaka, Y., Murakami, T., Nishimura, J., Yoshida, A. & Fenimore, E. E. *Astrophys. J.* (in the press).
6. Mazets, E. et al. *Astrophys. Space Sci.* **80**, 1–143 (1982).
7. Matz, S. et al. in *Proc. Los Alamos Workshop on Gamma Ray Bursts* (eds Ho, C., Epstein, R. & Fenimore, E. E.) (Cambridge University Press, in the press).
8. Chuang, K. W., White, R. S., Klebesadel, R. W. & Laros, J. G. *Astrophys. J.* (in the press).
9. Hartmann, D. & Epstein, R. I. *Astrophys. J.* **346**, 960–966, (1989).
10. Brock, M. N. et al. in *Proc. Gamma Ray Bursts 1991* (eds Fishman, G. J. & Paciesas, W. S.) (in the press).
11. Cline, T. L. et al. in *Proc. Gamma Ray Bursts 1991* (eds Fishman, G. J. & Paciesas, W. S.) (in the press).
12. Hurley, K. et al. in *Proc. GRO Annapolis Workshop* (in the press).
13. Schoenfelder, V. et al. *IAU Circ. No.* 5369 (1991).
14. Paczynski, B. *Astrophys. J.* **348**, 485–494, (1990).
15. Brainerd, J. J. *Nature* (in the press).
16. Atteia, J.-L. & Hurley, K. *Adv. Space Res.* **6**(4), 39–43, (1986).
17. Paczynski, B. *Acta Astr.* **41**, 157–166 (1991).
18. Paczynski, B. *Acta Astr.* **41**, (in the press).

ACKNOWLEDGEMENTS. We thank the members of the BATSE data operations team for processing the instrument data, and M. Finger for Monte Carlo calculations of burst distribution moments.

A planetary system around the millisecond pulsar PSR1257+12

A. Wolszczan* & D. A. Frail†

* National Astronomy and Ionosphere Center, Arecibo Observatory, Arecibo, Puerto Rico 00613, USA

† National Radio Astronomy Observatory, Socorro, New Mexico 87801, USA

MILLISECOND radio pulsars, which are old ($\sim 10^9$ yr), rapidly rotating neutron stars believed to be spun up by accretion of matter from their stellar companions, are usually found in binary systems with other degenerate stars¹. Using the 305-m Arecibo radiotelescope to make precise timing measurements of pulses from the recently discovered 6.2-ms pulsar PSR1257+12 (ref. 2), we demonstrate that, rather than being associated with a stellar object, the pulsar is orbited by two or more planet-sized bodies. The planets detected so far have masses of at least $2.8 M_{\oplus}$ and $3.4 M_{\oplus}$, where $M_{\oplus}$ is the mass of the Earth. Their respective distances from the pulsar are 0.47 AU and 0.36 AU, and they move in almost circular orbits with periods of 98.2 and 66.6 days. Observations indicate that at least one more planet may be present in this system. The detection of a planetary system around a nearby (~ 500 pc), old neutron star, together with the recent report on a planetary companion to the pulsar PSR1829–10 (ref. 3) raises the tantalizing possibility that a non-negligible fraction of neutron stars observable as radio pulsars may be orbited by planet-like bodies.

The 6.2-ms pulsar PSR1257+12 (Fig. 1) was discovered during the search at high galactic latitudes for millisecond pulsars conducted in February 1990 with the 305-m Arecibo radiotelescope at a frequency of 430 MHz (ref. 2). The characteristics of this survey and the details of data analysis are described else-

where⁴. The confirming observations made on 5 July 1990 have been followed by routine pulse timing measurements of the new pulsar. A total of 4,040 pulse time-of-arrival (TOA) observations have been accumulated so far, with the Arecibo radiotelescope, the 40-MHz, three-level correlation spectrometer and the Princeton Mark III pulsar processor at 430 MHz and 1,400 MHz. A typical uncertainty in the TOAs derived from 1-min pulse integrations is ~ 15 μ s.

The standard analysis of the timing data has been carried out using the model fitting program TEMPO⁵ and the Center for Astrophysics Solar System ephemeris PEP740R. With a growing time span of the TOA measurements, it had gradually become clear that the TOAs showed an unusual variability superimposed on an annual sinusoidal pattern caused by a small ($\sim 1'$) error in the assumed pulsar position. To separate these effects unambiguously, a timing-independent, interferometric position of PSR1257+12 was measured with the Very Large Array (VLA) in its A-array configuration on 19 July and again on 18 September 1991. The $\sim 0.1''$ accuracy of the resulting pulsar position was achieved by referencing the fringe phase to a point-source calibrator 1.7° away.

A least-squares fit of a simple model, which involved the pulsar's rotational period, P , and its derivative, $\dot{P}$, as free parameters and the fixed VLA position (Table 1), to the timing data spanning the period of 486 days resulted in post-fit residuals shown in Fig. 2a. The residuals, which measure the difference between the predicted and the actual TOAs, show a quasiperiodic 'wandering' over the entire pulsar period. A closer examination of this effect has revealed that it was caused by two strict periodicities of 66.6 and 98.2 days in the pulse arrival times. This is further demonstrated in Fig. 2b and c, which shows post-fit residuals after fitting each of the two periods separately to the above data, assuming simple keplerian binary models involving a low-mass binary companion to the pulsar. Evidently, fitting for one of the assumed binary periods leaves the other one as a post-fit residual, implying that the pulse arrival times of PSR1257+12 are indeed affected by two independent periodicities. Further detailed analysis has shown that the periodicities are independent of radio frequency and that other millisecond pulsars routinely observed at Arecibo with the same data acquisition equipment show no such effect in their timing residuals.

Millisecond pulsars are extremely stable rotators. Systematic timing observations of objects like the 1.5-ms pulsar 1937+21 (ref. 6) have not revealed any timing noise, quasiperiodic TOA variations or 'glitches' at the level often found in the population of younger pulsars and believed to be related to neutron star seismology⁷. The frequency independence of the amplitude of

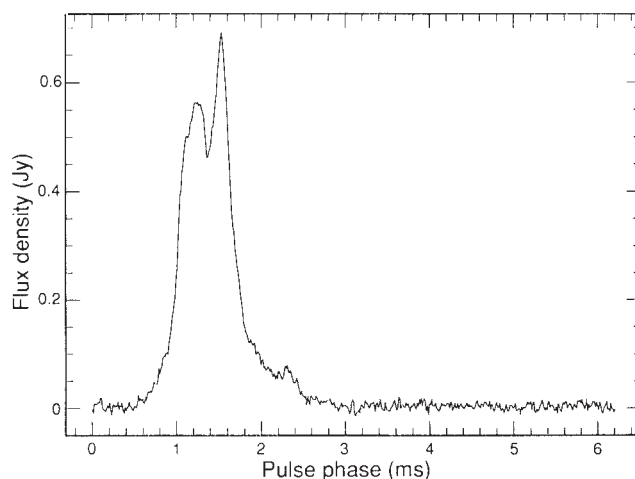


FIG. 1. The average pulse profile of PSR1257+12 at 430 MHz. The effective time resolution is ~ 12 μ s.

TABLE 1 Parameters of the PSR1257+12 system

Pulsar parameters			
Rotational period, P	$0.00621853193177 \pm 0.00000000000001$ s		
Period derivative, $\dot{P}$	$1.21 \times 10^{-19} \pm 2 \times 10^{-21}$ s s $^{-1}$		
Right ascension (B1950.0, VLA)	12 h 57 min 33.131 s ± 0.015		
Right ascension (B1950.0, timing)	12 h 57 min 33.126 s ± 0.003		
Declination (B1950.0, VLA)	$12^\circ 57' 05.9'' \pm 0.1$		
Declination (B1950.0, timing)	$12^\circ 57' 06.60'' \pm 0.02$		
Epoch	JD 2448088.9		
Dispersion measure	10.18 ± 0.01 pc cm $^{-3}$		
Flux density (430 MHz)	20 ± 5 mJy		
Flux density (1,400 MHz)	1.0 ± 0.2 mJy		
Surface magnetic field, B	8.8×10^8 G		
Characteristic age, τ_c	0.8×10^8 yr		
Keplerian orbital parameters			
Projected semimajor axis, $a_1 \sin i$	1.31 ± 0.01 light ms	1.41 ± 0.01 light ms	
Eccentricity, e	0.022 ± 0.007	0.020 ± 0.006	
Epoch of periastron, T_0	JD 2448105.3 ± 1.0	JD 2447998.6 ± 1.0	
Orbital period, P_b	5751011.0 ± 800.0 s	8487388.0 ± 1800.0 s	
Longitude of periastron, ω	$252^\circ \pm 20^\circ$	$107^\circ \pm 20^\circ$	
Parameters of the planetary system			
Planet mass, $m_{2,3}$ ($M_\oplus$)	$3.4/\sin i$	$2.8/\sin i$	
Distance from the pulsar, d (AU)	0.36	0.47	
Orbital period, P_b (days)	66.6	98.2	

TOA variations in PSR1257+12 rules out propagation phenomena, such as those detected in eclipsing binary pulsars^{8,9}, as a possible source of the observed periodicities. The integrated pulse profiles of PSR1257+12 do not show any morphological changes that might indicate the presence of a free precession of the pulsar spin axis^{10,11} or any magnetospheric phenomena at the level that could lead to periodic TOA variations of the measured magnitude. Consequently, the most plausible remaining alternative is that PSR1257+12 has two low-mass companions and that their orbital motion is responsible for the observed TOA variations.

To analyse this exciting possibility further, we have modified the code of the TEMPO program to accommodate a timing model including multiple, noninteracting keplerian orbits. The result of a fit of the model including the rotational and positional parameters of the pulsar as well as the two, five-parameter keplerian orbits to the entire data set is shown in Fig. 2d. The post-fit r.m.s. residual of the resultant timing model is $\sim 18 \mu\text{s}$ (comparable to the individual TOA uncertainties) and the remaining residuals contain very little systematic variation. The model parameters of the pulsar and its assumed two companions are listed in Table 1. The derived parameters have been obtained from standard considerations involving very low-mass objects in keplerian orbits around a $1.4M_\odot$ neutron star. The pulsar period variations predicted by this model are displayed in Fig. 3 together with the observed period values derived from timing data. Note that the presence of the planet-sized bodies orbiting the pulsar results in apparent period variations of only ± 15 ps. This is caused by the pulsar's spatial motions which are characterized by the projected maximum velocity and displacement amplitudes of $\pm 0.7 \text{ ms}^{-1}$ and $\pm 900 \text{ km}$, respectively.

The high quality of the two-companion model fit to the pulsar timing data covering several orbital cycles provides compelling evidence that PSR1257+12 possesses a planetary system consisting of at least two planet-sized bodies. The possibility of more planets around PSR1257+12 is indicated by a $0.7''$ discrepancy between the VLA and the timing positions, which is considerably greater than the conservative error estimates (Table 1). This discrepancy may arise from a bias in the timing position caused by a presence of a third planet with orbital period close to one year. Because the currently measured value of the second period derivative, $\dot{P}$, of the pulsar is not significant, the effect of any outer planets that could be present in the 1257+12 system would be entirely absorbed in $\dot{P}$.

The characteristics of the 1257+12 planets are not unlike those of the inner Solar System. Both planets circle the pulsar at distances similar to that of Mercury in its orbit around the Sun. Assuming a random distribution of the inclinations, i , of

orbital planes, there is a 50% probability that $i \geq 60^\circ$, so that the median values of the permissible masses are $3.2M_\oplus$ and $3.9M_\oplus$, respectively (see Table 1). Interestingly, the ratio of orbital periods, 1.476 ± 0.001 , is close to a 3/2 orbital resonance of the type often encountered in the Solar System, either between planetary satellites, between Jupiter and some asteroids, or even Neptune and Pluto¹². Also, the similarity of the measured eccentricities combined with the $\sim 180^\circ$ separation of the pericentres of the orbits can be easily understood in terms of secular perturbations of the orbital elements of the two planets (see, for example ref. 13). Finally, a hypothetical energy flux at the planets' distance from the pulsar ($\sim 0.4 \text{ AU}$) due to the neutron star's rotational energy loss ($\dot{E} = I\omega\dot{\omega}$, where I is the moment of inertia and $\omega = 2\pi/P$) is $\sim 4 \times 10^7 \text{ erg cm}^{-2} \text{ s}^{-1}$. This is ~ 30 times the solar constant and corresponds to a black-body temperature of $\sim 670 \text{ K}$, which is similar to the measured dayside surface temperature of Mercury (see, for example, ref. 14).

The possibility of the existence of planet-sized bodies orbiting neutron stars has been contemplated in the past^{15,16}. The most recent evidence has been presented by Bailes *et al.*³, who have detected a ~ 6 -month periodicity in the timing residuals of a relatively young pulsar, PSR1829-10, which is interpreted in terms of the orbital motion of a $\sim 10 M_\oplus$ companion. A number of mechanisms have been proposed to explain a planet around PSR1829-10 (refs 17-21). In the case of PSR1257+12, its old age and a low surface magnetic field (Table 1) are typical of neutron stars which are believed to evolve in low-mass binary systems and are spun up to the observed millisecond periods by accretion of matter from their stellar companions¹. As it is not very likely that any primordial planets would survive this

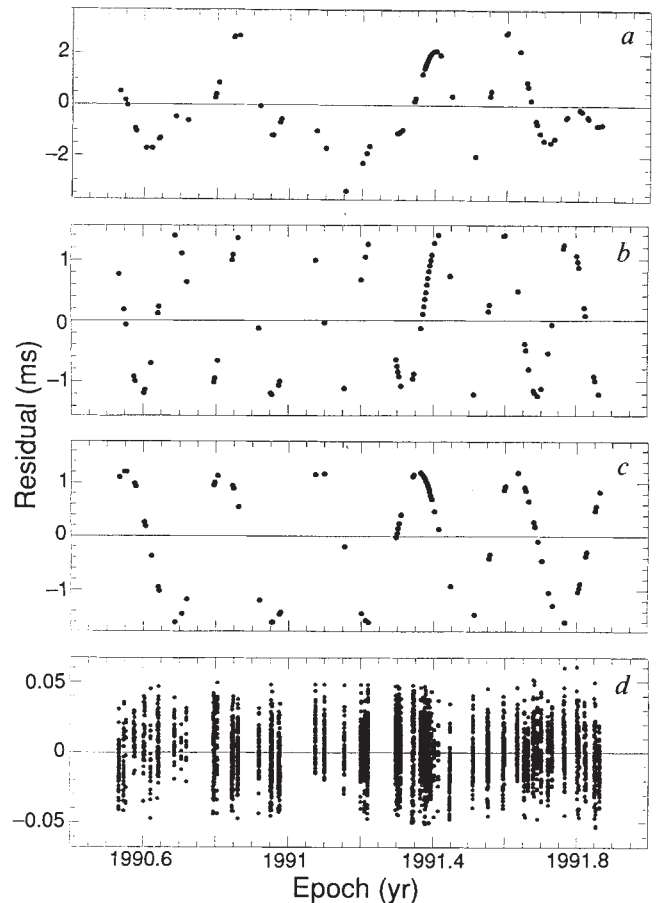


FIG. 2 The post-fit residuals of pulse arrival times from PSR1257+12. *a*, Fit for rotational parameters only (VLA position of the pulsar fixed); *b*, fit for a 98.2-day keplerian orbit (leaves a 66.6-day periodicity as residual); *c*, fit for a 66.6-day keplerian orbit (leaves a 98.2-day periodicity as residual); *d*, fit for all parameters of *a-c*.

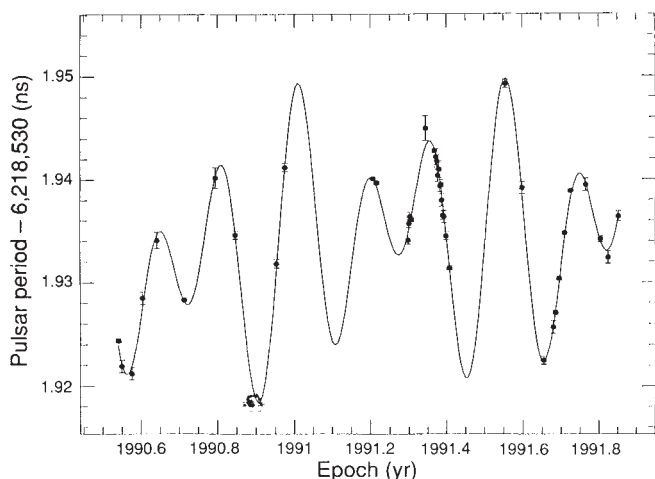


FIG. 3 Period variations of PSR1257+12. Each period measurement is based on observations made on at least two consecutive days. The solid line denotes changes in period predicted by a two-planet model of the 1257+12 system.

kind of evolution²², the observed 1257+12 system probably consists of 'second generation' planets created at or after the end of the pulsar's binary history. Because a second supernova explosion is not expected to occur in a low-mass binary, such planets could form a stable system of bodies orbiting an old neutron star. Another important evolutionary constraint is provided by the observed very low eccentricities ($e \approx 0.02$) of the planetary orbits and the near-resonance ratio of the orbital periods. These characteristics suggest that the planets were created from some form of accretion disk that would naturally provide means to circularize the orbits and to bring them close to a 3/2 resonance¹². Consequently, it seems that any plausible mechanism for the creation of a planetary system around a millisecond pulsar must provide a way to remove its stellar companion in a manner different from the disruption of a binary system caused by a supernova explosion and to retain enough circumpulsar matter to form planet-sized objects.

Within the constraints provided by the observational evidence, it is tempting to postulate that the 1257+12 system simply represents one of the possible outcomes of a neutron star evolution in a low-mass binary system¹. As two examples of companion-star evaporation by millisecond pulsars have already been detected^{7,8}, this mechanism seems to account naturally for the absence of such a companion to PSR1257+12. In this case, one way to supply the material for planet creation could be a disk formed out of the ablated stellar matter ejected from the binary system^{23,24}. The existence of such an 'outer disk' around the eclipsing binary pulsar PSR1957+20 could explain the observed orbital decay of this system²⁵.

The results described here strongly suggest that one of the nearby galactic millisecond pulsars, PSR1257+12, is accompanied by a system of two or more planet-sized bodies. The pulse timing method (similar to the analysis of single-line spectroscopic binaries) was used to detect a $\pm 0.7 \text{ m s}^{-1}$ pulsar 'wobble' caused by orbital motions of the planets. In fact, planetary masses smaller than that of the Moon could easily be detected with this technique. At present, this kind of accuracy is entirely inaccessible to the optical methods of planetary system detection²⁶. The existence of both the 1257+12 planetary system and the object orbiting the pulsar PSR1829-10 seems to suggest that planets can form under a variety of conditions. This notion and the possibility of a non-negligible frequency of occurrence of planets around neutron stars, if confirmed, will have far-reaching consequences for our understanding of the formation and evolution of planetary systems and for future strategies of searches for planets outside the Solar System. □

Received 21 November; accepted 9 December 1991.

1. Bhattacharya, D. & van den Heuvel, E. P. J. *Phys. Rep.* **203**, 1-124 (1991).
2. Wolszczan, A. *IAU Circ. No.* 5073 (1990).
3. Bailes, M., Lyne, A. G. & Shemar, S. L. *Nature* **352**, 311-313 (1991).
4. Wolszczan, A. *Nature* **350**, 688-690 (1991).
5. Taylor, J. H. & Weisberg, J. M. *Astrophys. J.* **345**, 434-450 (1989).
6. Davis, M. M., Taylor, J. H., Weisberg, J. M. & Backer, D. C. *Nature* **315**, 547-550 (1985).
7. Alpar, M. A., Nandkumar, R. & Pines, D. *Astrophys. J.* **311**, 197-213 (1986).
8. Fruchter, A. S., Stinebring, D. R. & Taylor, J. H. *Nature* **333**, 237-239 (1988).
9. Lyne, A. G. et al. *Nature* **347**, 650-652 (1990).
10. Shaham, J. *Astrophys. J.* **214**, 251-260 (1977).
11. Nelson, R. W., Finn, L. S. & Wasserman, I. *Astrophys. J.* **348**, 226-231 (1990).
12. Peale, S. J. *Ann. Rev. Astr. Astrophys.* **14**, 215-246 (1976).
13. Dermott, S. F. & Nicholson, P. D. *Nature* **319**, 115-120 (1986).
14. Vilas, F. in *Mercury* (eds Vilas, F., Chapman, C. R. & Matthews, M. S.) 59-76 (University of Arizona Press, Tucson, 1988).
15. Richards, D. W., Pettengill, G. H., Counselman, C. C. III & Rankin, J. M. *Astrophys. J.* **160**, L1-6 (1970).
16. Demiański, M. & Prószyński, M. *Nature* **282**, 383-385 (1979).
17. Podsiadlowski, Ph., Pringle, J. E. & Rees, M. J. *Nature* **352**, 783-784 (1991).
18. Fabian, A. C. & Podsiadlowski, Ph. *Nature* **353**, 801-801 (1991).
19. Lin, D. N. C., Woosley, S. E. & Bodenheimer, P. H. *Nature* **353**, 827-829 (1991).
20. Krolik, J. H. *Nature* **353**, 829-831 (1991).
21. Wasserman, I., Cordes, J. M., Finn, L. S. & Nelson, R. W. *Cornell Univ. preprint* (1991).
22. Nakano, T. *Mon. Not. R. astr. Soc.* **224**, 107-130 (1987).
23. Shu, F. H., Lubow, S. H. & Anderson, L. *Astrophys. J.* **29**, 223-241 (1979).
24. Rudak, B. & Paczyński, B. *Acta Astr.* **31**, 13-24 (1981).
25. Ryba, M. F. & Taylor, J. H. *Astrophys. J.* **380**, 557-563 (1991).
26. Black, D. C. *Space Sci. Rev.* **25**, 35-81 (1980).

ACKNOWLEDGEMENTS. We thank D. Backer, P. Nicholson, B. Paczyński, F. Rasio, S. Shapiro, J. Taylor and S. Teukolsky for discussions. We also thank J. Taylor for making available the Princeton Mark III processor. Arecibo Observatory is part of the National Astronomy and Ionosphere Center, which is operated by Cornell University under contract with the NSF. The VLA is operated by the NRAO under cooperative agreement with the NSF. D.A.F. is supported by an NSERC postdoctoral fellowship and an NRAO Jansky fellowship.

Birth of a radio supernova remnant in supernova 1987A

L. Staveley-Smith*, R. N. Manchester*,
M. J. Kesteven*, D. Campbell-Wilson†,
D. F. Crawford†, A. J. Turtle†, J. E. Reynolds*,
A. K. Tzioumis*, N. E. B. Killeen* & D. L. Jauncey*

* Australia Telescope National Facility, CSIRO, PO Box 76,
New South Wales, 2121, Australia

† School of Physics, University of Sydney, New South Wales 2006,
Australia

FOLLOWING an initial radio outburst¹ which decayed on a time-scale of a few weeks, supernova 1987A² was undetectable at radio wavelengths until recently. In mid-1990, 1,200 days after the explosion, radio emission was once again detected³ with the Molonglo Observatory Synthesis Telescope and the Australia Telescope Compact Array. Our observations since then show that the source has increased in strength at all monitored frequencies between 843 MHz and 8.6 GHz, with considerable variations in spectral index. The extended radio emission is centred within 0.5 arcsec of the optical supernova, and probably lies within the [O III] ring imaged by the Hubble Space Telescope. We interpret the emission as optically thin synchrotron emission from shock-accelerated electrons. This is the first time the birth of a nearby radio supernova remnant has been witnessed, and future observations will allow the structure of the remnant to be compared with the many other known radio remnants.

Since the demise of the initial radio emission, we have monitored SN1987A using the Molonglo Observatory Synthesis Telescope⁴ (MOST) at 843 MHz and, since mid-1989, the Australia Telescope Compact Array⁵ (ATCA) at 1.5, 2.4, 4.8 and 8.6 GHz. Renewed radio emission at the supernova position was first detected on 6 July 1990, using the MOST³. The flux density at this time was $\sim 3 \text{ mJy}$. At higher frequencies, the onset of the radio emission was delayed, the first detection with the ATCA at 4.8 GHz being on 16 August 1990 (0.7 mJy). Since then, the MOST and the ATCA have made 70 and 40 observations, respectively. Each of these observations lasted ~ 12 hours. The MOST observations are affected by the strong source 30

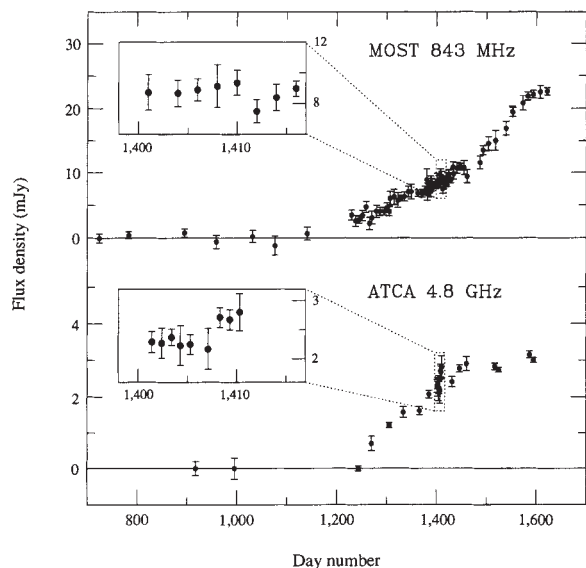


FIG. 1 Flux density against days since the geocentric time of the supernova explosion for the supernova remnant 1987A at two frequencies: 843 MHz (Molonglo Observatory Synthesis Telescope) and 4.8 GHz (Australia Telescope Compact Array). The insets are expanded plots centred on day 1,406.

Doradus, 20 arcmin away. Careful calibration has reduced the final r.m.s. flux uncertainty to 0.9 mJy. The zero level is determined from the mean of 11 observations in 1988 and 1989 when the supernova flux density is believed to be less than 0.5 mJy at 843 MHz. We made ATCA observations using a maximum baseline of ~ 3 km up to the end of March 1991 and 6 km thereafter. Because of the smaller primary beam and the reduced sensitivity to extended emission, the ATCA observations are not significantly affected by the 30 Dor region. Flux densities for the source at the supernova position (Fig. 1) show that the source has increased in strength at both frequencies. The time variation is, however, different at the two frequencies. At 843 MHz a linear increase is followed, at about day 1,470 (after the geocentric date of the supernova explosion), by a steeper rise which may now be flattening, whereas at 4.8 GHz an initially linear increase flattened out at about day 1,450.

To determine the minimum timescale for variations in the radio emission, frequent observations were made using both the MOST and the ATCA for a 2-week period centred on 30 December 1990. These observations are shown on an expanded scale in the insets to Fig. 1. At 4.8 GHz there is tentative evidence for an increase in flux density on day 1,408. Between days 1,401 and 1,407, the flux density was constant within the uncertainties at 2.26 ± 0.07 mJy, whereas from days 1,408 to 1,410 it rose to 2.73 ± 0.07 mJy, an increase of $21 \pm 5\%$. During this period, there was no significant change in the measured flux density of a 35-mJy background source located within the observed field ~ 4 arcmin southeast of the supernova. The flux density jump was not reflected at 843 MHz. A flux density increase of $\sim 45\%$ at 843 MHz was, however, noted about day 1,305. Although this increase was not as well sampled in time, it gives additional evidence for variability in the radio source on timescales shorter than the 50–100 days indicated by the general upward trend in flux density. The upper limits to the polarized flux density at 4.8 GHz between days 1,408 and 1,410 are 10% for the integrated linear and 7% for the integrated circular polarization.

The long-term flux density variations apparent in Fig. 1 indicate considerable changes in the spectral index of the radio emission. In Fig. 2 we show spectra obtained at four different epochs starting on day 1,243 (20 July 1990, UT). These include observations made with the ATCA at 1.5, 2.4 and 8.6 GHz as well as the 843-MHz and 4.8-GHz observations above. All observations were made within two days of the nominal day number. Initially, the radio spectrum was steep, with a spectral index α ($S \propto \nu^\alpha$) less than -1.6 (95% confidence) between 843 MHz and 4.8 GHz. As the source intensity increased, the spectrum flattened, reaching $\alpha = -0.85 \pm 0.02$ at about day 1,386.

At the time of writing, the spectrum appears to be steepening again, with $\alpha = -1.08 \pm 0.04$ at day 1,595.

An observation of duration 4 hours on 10 December 1990 made with the Parkes–Tidbinbilla interferometer⁶ at a frequency of 2.3 GHz showed no evidence for a compact component associated with SN1987A. An upper limit (5σ) of 1.2 mJy was obtained. A simultaneous ATCA observation gave a 2.3-GHz flux density of 3.9 ± 0.4 mJy for the supernova. This shows that the angular extent of the source is greater than 0.1 arcsec and excludes a compact source, such as a central pulsar, as a principal source of emission. In addition, the steep spectral index of the extended emission argues against a pulsar-powered

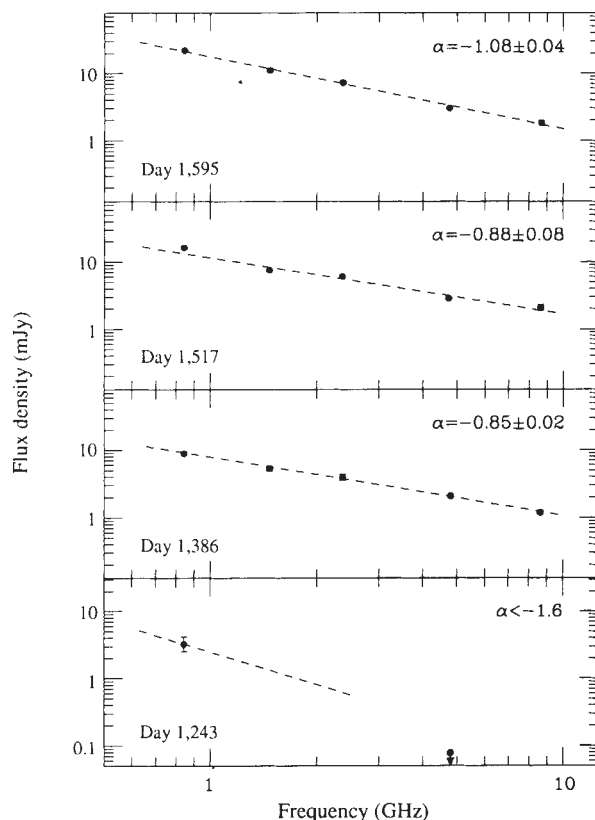


FIG. 2 The spectral evolution of SNR1987A at radio wavelengths. The initial spectrum (day 1,243) was steep, followed by a flattening to $\alpha = -0.8$, then steepening again to $\alpha = -1.08 \pm 0.04$ at day 1,595.

remnant, which would probably have a flatter ($\alpha > -0.5$) spectral index than that observed.

Because of the possibility of using light travel time to map out the geometry of the emitting region, the commissioning of the ATCA 6-km antenna was brought forward to April 1991. Two 12-hour observations made using the 6-km array at 8.6 GHz on 25 June and 6 July 1991 have been combined with the Hubble Space Telescope [O III] image⁷ of SN1987A to produce the image shown in Fig. 3. The centroid of the radio emission, obtained using the nearby source PKS 0530-727 as a position calibrator, is at RA 05 h 35 min 28.00 s, Dec. $-69^{\circ}16'11.1''$ (J2000) with formal errors of 0.05 arcsec in each coordinate. This position is 0.5 arcsec north of the reported optical supernova position⁸ of RA 05 h 35 min 28.00 s, Dec. $-69^{\circ}16'11.6''$ (J2000). The principal source of uncertainty in the position of the radio source with respect to the optical image arises from uncertainties in the registration of the optical and radio reference frames. A comparison of accurate optical and radio positions⁹ of two nearby extragalactic sources, PKS 0637-752 and PKS 0743-673, shows no significant offset in the reference frames, but the errors in each coordinate are ~ 0.15 arcsec. Combined with the optical and radio position errors (~ 0.1 arcsec), the total optical-radio position uncertainty is ~ 0.2 arcsec. Therefore, although the nominal radio centroid lies closer to the northern blueshifted¹⁰ side of the [O III] ring, it is possible that the radio source is actually centred on the supernova position. With the 0.9-arcsec resolution of the 8.6-GHz image, the radio source is seen to be extended, but with the bulk of the emission coming from a region smaller in area than that enclosed by the [O III] ring surrounding the supernova. The deconvolved size of the radio source is $(1.0 \pm 0.1) \times (0.9 \pm 0.1)$ arcsec at position angle $\sim 110^{\circ}$, compared with the [O III] ring dimensions⁷ of 1.7×1.2 arcsec at position angle $\sim 80^{\circ}$.

These observations demonstrate that we are witnessing the birth of a radio supernova remnant, SNR1987A. The spectra are consistent with optically thin synchrotron radiation; the

resolution of the source by the Parkes-Tidbinbilla interferometer and ATCA shows that the emission region is extended. We propose, therefore, that the emission is being generated by shock excitation of relativistic electrons as the supernova blast wave collides with clumps of circumstellar material. The observed overall size of the radio source combined with the uncertainty in the registration with the Hubble Space Telescope image strongly suggest that this material lies between the supernova and the [O III] ring. This hypothesis is consistent with the absence so far of soft X-ray emission from the region; strong X-ray emission is expected when the blast wave encounters the relatively dense ring material¹¹. The size implies a velocity for the blast wave of $\sim 30,000$ km s⁻¹, a value which is consistent with optical and radio estimates^{12,13} of the highest ejecta velocities, and suggests that the first observed impact with the ring will be in early 1993.

The delay in onset of the higher-frequency radio emission relative to that at 843 MHz, and the associated increase in radio spectral index with time until about day 1,500, show that lower-energy relativistic electrons are generated well before those at higher energies. Because of the unknown geometry of the emission region, it is difficult to disentangle light travel-time effects from timescales for acceleration of the electrons. But it is clear from the initial switch-on, and the recent surge in the low-frequency flux density, that the acceleration timescale may be up to a few months. The rapid time variation observed at 4.8 GHz around day 1,408 and at 843 MHz around day 1,305, suggests that some electrons may be more efficiently accelerated. The present volume of the synchrotron-emitting region is $\sim 10^{47} \phi$ m³, where ϕ is the volume filling factor. For the extrapolated 10 MHz-10 GHz radio luminosity of 3×10^{25} W, the minimum kinetic and magnetic energy requirement¹⁴ is $4 \times 10^{39} \phi^{3/7}$ J for a relativistic electron-to-proton energy ratio of 0.01. Filling factors below unity are obtained only for equipartition field strengths in excess of 3×10^{-7} T (3 mG). This suggests that a considerable volume of the remnant is filled with radiating particles and sizeable magnetic fields, or that particle energy density is far from equilibrium with the magnetic energy density, or that the ionic energy density is negligible. To explain the very small filling factors implied by rapid time variability, some combination of the above may be necessary.

The recent rapid increase at 843 MHz and the flattening at higher frequencies indicate that the blast wave has reached a new clump of circumstellar gas. The 4.8-GHz emission may resume its increase within the next few months as electrons in the new region are accelerated to higher energy. Future observations will help to establish parameters of the electron acceleration process and the relationship between the radio and optical emitting regions. □

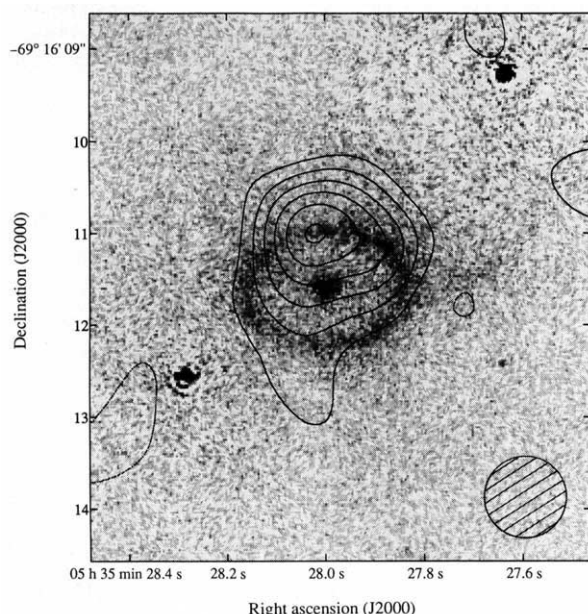


FIG. 3 Contours of 8.6-GHz radio emission (Australia Telescope) overlaid on the Hubble Space Telescope [O III] image⁷. The resolution of the radio image is 0.9 arcsec, as indicated by the beam plotted in the lower right; the resolution of the Hubble image is 0.07 arcsec. The Hubble image has no accurate position calibration, and therefore has been shifted so that the supernova lies at the mean position quoted in ref. 8. Some or all of the apparent offset between the centroid of the radio emission and the supernova (at the centre of the ring) may be due to misalignment of the optical and radio reference frames. Contour levels are $-0.15, 0.15, 0.3, 0.45, 0.6, 0.75$ and 0.9 mJy per beam.

Received 9 September; accepted 30 October 1991.

1. Turtle, A. J. *et al. Nature* **327**, 38-40 (1987).
2. Shelton, I. *IAU Circ.* 4316 (1987).
3. Turtle, A. J., Campbell-Wilson, D., Manchester, R. N., Staveley-Smith, L. & Kesteven, M. J. *IAU Circ.* 5086 (1990).
4. Mills, B. Y. *Proc. astr. Soc. Austr.* **4**, 156-159 (1981).
5. Norris, R. P. *Sky Telescope* **76**, 615-620 (1988).
6. Norris, R. P., Kesteven, M. J., Wellington, K. J. & Batty, M. J. *Astrophys. J. Suppl. Ser.* **67**, 85-91 (1988).
7. Jakobsen, P. *et al. Astrophys. J.* **369**, L63-66 (1991).
8. White, G. L. & Malin, D. *ESO Workshop on the SN1987A*, 11-17 (European Southern Observatory, Garching, 1987).
9. Harvey, B. R. *et al. Astr. J.* (in the press).
10. Wood, P. R. & Faulkner, D. J. *IAU Circ.* 4739 (1989).
11. Luo, D. & McCray, R. *Astrophys. J.* **372**, 194-198 (1991).
12. Hanuschik, R. W. & Dachs, J. *Astr. Astrophys.* **182**, L29-30 (1987).
13. Jauncey, D. L. *et al. Nature* **334**, 412-415 (1988).
14. Pacholczyk, A. G. *Radio Astrophysics* 171 (Freeman, San Francisco, 1970).

ACKNOWLEDGEMENTS. We thank J. Lim, D. Berinson, M. O'Connor, Qiao Guojun, I. Collings and M. Bulmer for assistance with the ATCA observations; R. Wark, R. Ferris and R. Norris for assistance with the PTI observations; R. Sault for assistance with the analysis of the polarization data; and L. Ball, D. Melrose, G. Bicknell and R. Ekers for discussions. We especially thank the ATNF staff for the accelerated commissioning of the 6-km antenna. The Australia Telescope National Facility is operated in association with the Division of Radiophysics by CSIRO. The MOST is supported by the Australian Research Council and Science Foundation for Physics within the University of Sydney.

Photochemical bromine production implicated in Arctic boundary-layer ozone depletion

J. C. McConnell*, G. S. Henderson†, L. Barrie‡,
J. Bottenheim‡, H. Niki§, C. H. Langford||
& E. M. J. Templeton*

* Department of Earth and Atmospheric Sciences, York University,
4700 Keele Street, North York, Ontario, Canada M3J 1P3

† Department of Geology, University of Toronto, 22 Russell Street,
Toronto, Ontario, Canada M5S 3B1

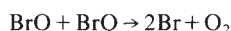
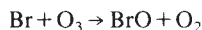
‡ Atmospheric Environment Service, 4905 Dufferin Street, North York,
Ontario, Canada, M3H 5T4

§ Department of Chemistry, York University, 4700 Keele Street,
North York, Ontario, Canada M3J 1P3

|| Department of Chemistry and Biochemistry, Concordia University,
Montréal, Québec, Canada H3G 1M8

RECENT measurements¹⁻⁷ in the Arctic have revealed episodic destruction of boundary-layer ozone from 30–40 parts per 10⁹ by volume (p.p.b.v.) to undetectable levels on a timescale of less than a day, during periods when the boundary layer is very stable. The ozone destruction begins at polar sunrise, continues for the months of March and April, and is strongly associated with levels of filterable bromine which are much greater than during the rest of the year. Here we suggest that sea-salt Br⁻ reaches high concentrations in the snow pack during the long polar night, and is evolved into the atmosphere as Br₂ at polar sunrise. Ordinarily, gas-phase photochemistry would convert Br₂ to HBr or brominated organic compounds with consequently little destruction of boundary-layer ozone. In view of several laboratory experiments⁸⁻¹¹, and by analogy with the marine boundary layer¹², we propose that the HBr and brominated organic compounds will be scavenged by the ambient aerosols and ice crystals, and that these heterogeneous reactions release Br₂ back to the atmosphere. We argue that this cycling of bromine between the aerosol and the gas phase should maintain sufficiently high levels of Br atoms and BrO radicals to destroy ozone, in agreement with observations¹⁻⁷.

Two questions must be addressed if gas-phase Br chemistry is identified as the origin of the O₃ depletion. The first is the source of the high levels of filterable Br (f-Br, the sum of Br on particles and gaseous species such as HBr efficiently collected by Whatman 41 cellulose acetate or a combination Teflon/nylon filter); the second is a means to recycle the HBr and brominated organic compounds (BOCs) generated to more chemically active forms of bromine. There are problems with previously suggested mechanisms involving bromoform (CHBr₃)¹ and nitryl bromide (BrNO₂)¹³; CHBr₃ photolysis is probably too slow to generate enough Br, and generation of enough BrNO₂ requires a longer residence time for Arctic air than is usually observed¹⁴. The observations of Barrie *et al.*¹ and Bottenheim *et al.*⁷ can be explained by a mechanism involving the conversion of sea-salt-derived Br⁻ to gas-phase Br₂(g) through photoactivated reactions on aerosols and the snow pack surface. Photolysis of the liberated Br₂ rapidly forms Br, which can then destroy O₃ through the catalytic cycle

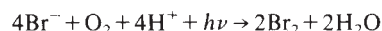


Gas-phase reactions of Br with species such as HO₂ and HCHO rapidly produce HBr and other BOCs which would normally shut down further depletion of O₃.

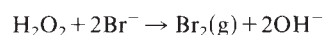
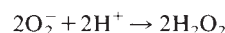
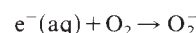
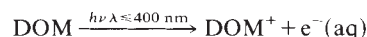
Particles (aerosols and ice crystals) are, however, ubiquitous in the Arctic boundary layer^{15,16} and are likely to be efficient scavengers of HBr and BOCs such as BrCHO, generated by reaction of Br with hydrocarbons (the aerosol-molecule collision frequency is ~10⁻²–10⁻³s⁻¹ for aerosol densities typical

of winter Arctic haze²). Marine boundary-layer measurements¹² suggest rapid recycling of Br⁻ from aerosol to gas phase during the day. We expect that the Br⁻ on the Arctic aerosol will be photoreleased and continually recycled from the particulate to the gas phase for the lifetime of the aerosol.

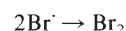
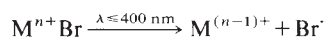
The mechanism we suggest requires photoinduced conversion of Br⁻ to gas-phase Br₂. The concept of photoconversion of halogens X⁻ → X₂(g) is not new and has most recently been proposed to explain the Cl deficit of marine aerosols¹⁷. Duce *et al.*¹⁸ have suggested that



is a possible conversion pathway. The detailed reaction sequence is unknown; two alternative pathways are⁸



(where DOM is dissolved organic matter), or⁹



(where M is a transition metal such as Fe(III), Mn(IV) or Cu(II)). We note the part played by H₂O₂, which is produced in abundance by the gas-phase chemistry. The gas-phase H₂O₂ could be readily scavenged by the aerosols and contribute to the maintenance of the Br₂(g) generation.

As for the source of the elevated bromine levels, sea spray is continually deposited on the snow pack throughout the polar night from open oceans at the periphery of the Arctic ice, and from leads and polynyas which apparently can occupy up to 12% of the Arctic ice area¹⁹. Additionally, sea salt can be moved by several mechanisms through the 3–4 m of crack-ridden sea ice into the snow pack. We are not aware of any measurements

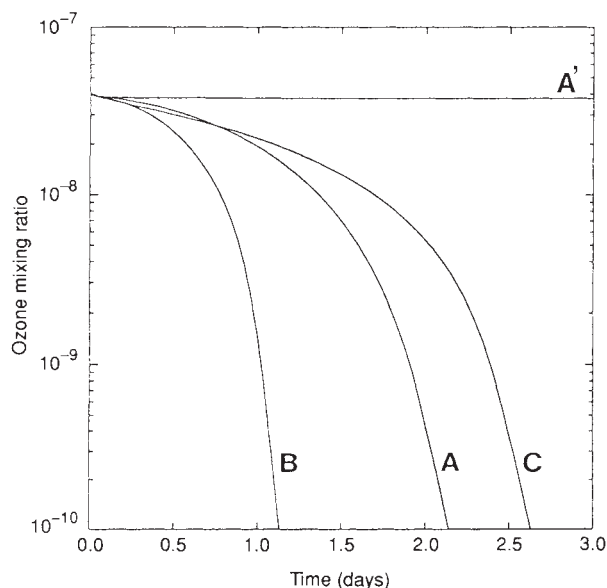
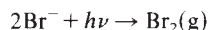


FIG. 1 Boundary-layer O₃ concentrations against time. Curve A is for standard initial conditions described in the text. Curve B is for the C₂H₂, C₂H₄ initial densities reduced by a factor of 4. Curve C shows the O₃ loss when the initial HCHO and CH₃CHO densities are set to zero. Curves A–C include simulation of heterogeneous chemistry. Curve A' is for the same conditions as A but with no heterogeneous mechanism operating.

of Br^- in the polar snow pack, but concentrations of Na^+ and Cl^- in the snow pack on the Arctic ocean (ref. 20, and R. M. Koerner, personal communication) are 100–600 times greater than those measured in fresh snow on the nearby Agassiz Ice Cap of Ellesmere Island. This would suggest a high probability of elevated Br^- levels in the ice-pack snow surface.

We have numerically simulated the impact that gas-phase Br will have on O_3 depletion. The gas-phase Br reaction scheme of Barrie *et al.*¹ has been augmented with the reactions of Br with CH_3CHO , C_2H_2 and C_2H_4 (as well as their products; ref. 21, and N. Peng *et al.*, manuscript in preparation) which have not been included in earlier calculations. In addition, we include reactions to simulate scavenging of Br compounds by aerosols with subsequent photoinduced heterogeneous conversion to Br_2



The reaction rates adopted for these heterogeneous processes are $\sim 10^{-4} \text{ s}^{-1}$, reflecting the rates required to duplicate measurements¹² of the diurnal behaviour of inorganic Br and aerosol Br.

For the numerical simulations we assume that the $\text{Br}_2(\text{g})$ has

already been released from the snow pack at polar sunrise and investigate the possibility that the cycling of the Br_x from aerosol to gas phase will lead to a substantial depletion of O_3 . The standard initial conditions adopted are based on the measurements reported by Bottenheim *et al.*⁷ and Kieser *et al.*²². We use 800 parts per 10^{12} (p.p.t.v.) C_2H_2 , 200 p.p.t.v. C_2H_4 , 40 p.p.t.v. HCHO , 60 p.p.t.v. CH_3CHO , and 50 p.p.t.v. NO_2 (typical of general background concentrations during the winter), respectively. Initial $\text{Br}_2(\text{g})$ levels are set at 50 p.p.t.v. (that is, 100 p.p.t.v. of Br_x), a value based on the measurements of Berg *et al.*²³ during periods of high f-Br in April at Barrow, Alaska. The simulations, for the most part, are performed for a 48-hr period with solar irradiance conditions typical of 82.5°N during the period around 1 April. The conditions outlined above, including the simulation of heterogeneous chemistry, represent scheme A: scheme A' is the same as A but without heterogeneous chemistry. Scheme B is for the above conditions but with C_2H_2 and C_2H_4 reduced by a factor of 4, to make them more typical of the values during the depletion episodes^{7,22}. Finally, scheme C is similar to that for B but with initial HCHO and CH_3CHO mixing ratios set at zero to simulate the possibility that the aldehydes measured by Bottenheim *et al.*⁷ are actually produced by the depletion mechanism.

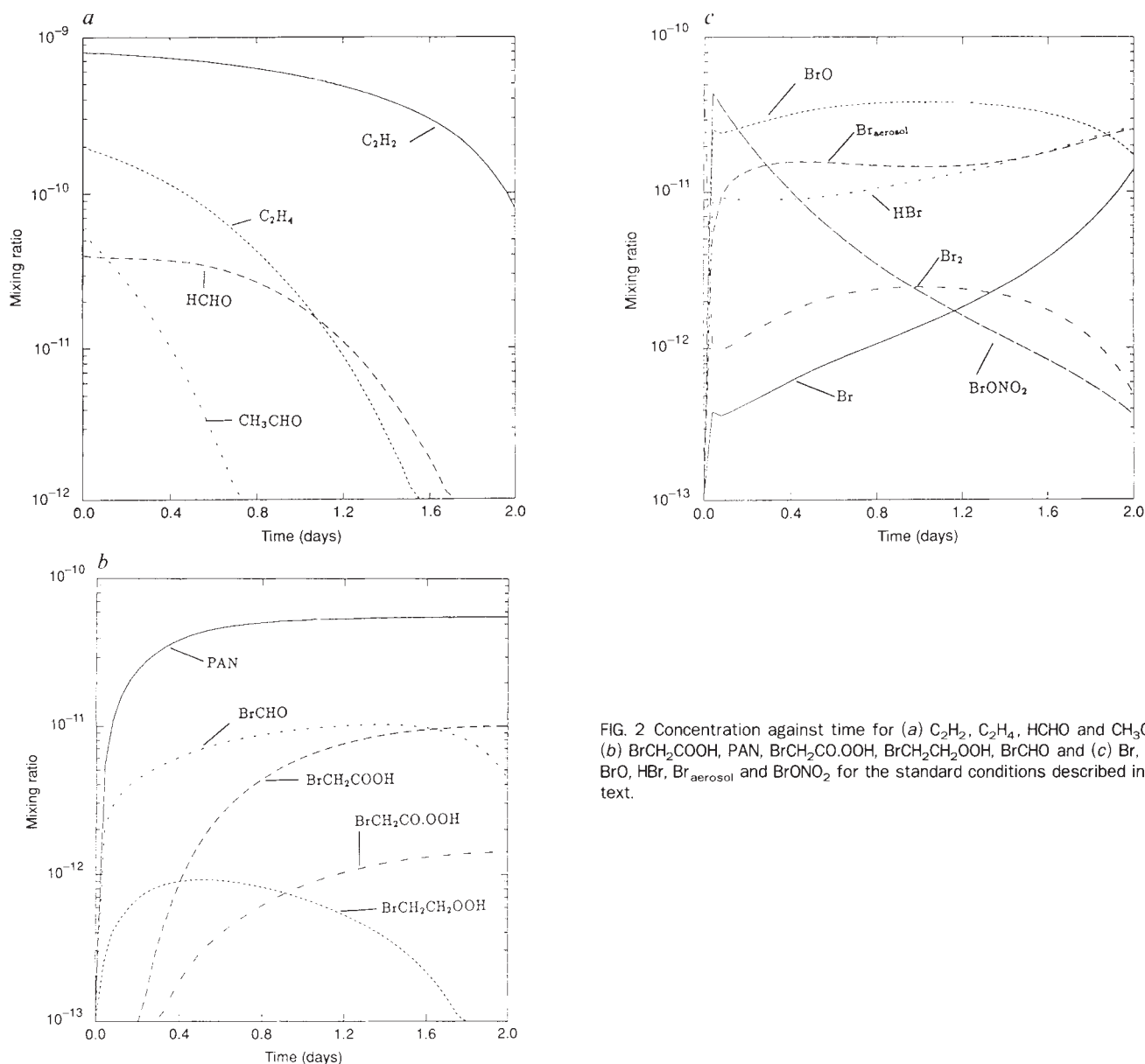


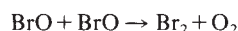
FIG. 2 Concentration against time for (a) C_2H_2 , C_2H_4 , HCHO and CH_3CHO ; (b) BrCH_2COOH , PAN , $\text{BrCH}_2\text{COOOH}$, $\text{BrCH}_2\text{CH}_2\text{OOH}$, BrCHO and (c) Br , Br_2 , BrO , HBr , $\text{Br}_{\text{aerosol}}$ and BrONO_2 for the standard conditions described in the text.

The results for scheme A (curve A of Fig. 1) show that O_3 is destroyed by the high levels of Br within one to two days. If the heterogeneous reactions are not active, then for standard conditions (curve A' of Fig. 1) the O_3 depletion after two days is no more than 10% because the Br_x rapidly ends up as HBr or BOCs, shutting down the destruction cycle, to give results similar to those of Barrie *et al.*¹ For scheme B, the depletion is even more rapid, with 90% depletion occurring in less than a day. With the CH_3CHO present the Br attack results in peroxyacetyl nitrate (PAN) formation which removes NO_2 and allows the active bromine to catalyse the O_3 destruction. If NO_2 is reduced from 50 to 10 p.p.t.v. in this scheme, the depletion is more rapid and occurs within a day. Without the CH_3CHO (scheme C of Fig. 1), the NO_2 ties up the active bromine as $BrONO_2$ so that O_3 depletion is substantially slowed but still complete after three days. (Of course it is possible that $BrONO_2$ itself, which we have not scavenged in the model, may in fact be scavenged by the aerosol, in which case scheme C would be more like B.)

One of the additional features of the proposed chemical scheme is the rapid destruction of C_2H_2 , C_2H_4 , HCHO and CH_3CHO (Fig. 2a) while generating BOCs (Fig. 2b). This may account in part for the lowered concentrations of C_2H_2 and C_2H_4 measured during the O_3 -depleted episodes at Alert⁷. Additionally, we find that with CH_3CHO present at measured concentrations to induce NO_2 'absorption' through the formation of PAN, the proposed mechanism remains rapid for Br levels as low as 45 p.p.t.v.

In our proposed scheme, HCHO and CH_3CHO act as enhancers for depletion. This is in contrast to the pure gas-phase mechanism outlined by Barrie *et al.*², where any aldehydes present act as inhibitors because of the production of HBr. The most critical aspect of our chemical mechanism is that HBr formation does not terminate the Br catalytic scheme, because $Br_2(g)$ is regenerated by heterogeneous photo-oxidation of Br^- in aerosols. Although the mechanism that we propose can account for many of the reported observations, such as the spring peak of f-Br and the rapid rate of the O_3 destruction, other chemical mechanisms are not excluded. For example, an alternative to gas-phase O_3 depletion could be heterogeneous destruction of O_3 directly on the photoactivated surface of the brominated aerosols and snow pack; in smog chamber experiments^{10,11} where O_3 has been absorbed on HCl-treated walls, $Cl_2(g)$ has been generated in both the dark and the light.

According to our hypothesis, more Br_2 will be driven off the snow as the solar flux becomes more intense, until either the snow is depleted or the boundary inversion layer finally breaks down, diluting the high levels of Br_x with free tropospheric air. Thus we would expect there to be a Br^- maximum in spring, as has been observed¹. The main forms of Br_x during the day are shown in Fig. 2c. In the early spring, however, in the narrow window before the onset of 24-hr sunlight, we would expect high concentrations of Br_2 to be formed when the Sun sets, through the reaction



During the night, we expect that HBr (and probably many of the BOCs) would be scavenged by the aerosols, whereas BrO would be converted to Br_2 , which would probably remain in the gas phase. Thus we expect a diurnal cycle for the gas-phase Br_x component of atmospheric Br. Sturges *et al.*⁶ do not seem to have observed such a cycle, but this may be because the boundary layer is less stable at Barrow and this influences the surface observations.

Given the putative presence of high concentrations of Br, BrO and brominated organics, it would seem prudent to determine whether the filters used to determine gas-phase inorganic Br and f-Br are efficient collectors of the species likely to be present, particularly Br_2 . Furthermore, as the ocean, ice leads, polynyas and snow pack are important for providing the Br source,

measurements of the spatial distribution of gas-phase inorganic bromine, non-methane hydrocarbons and Br compounds in surface snow would help in testing the proposed mechanism. Measurements of the Br^- concentrations with depth in the snow pack and polar sea ice would also help confirm the source of the Br^- levels. Finally, further laboratory experiments should be undertaken to confirm the details of the photosensitized Br_2 evolution from heterogeneous surfaces. □

Received 29 July; accepted 9 October 1991.

1. Barrie, L. A., Bottenheim, J. W., Schnell, R. C., Crutzen, P. J. & Rasmussen, R. A. *Nature* **334**, 138–141 (1988).
2. Barrie, L. A., den Hartog, G., Bottenheim, J. W. & Landsberger, S. J. *atmos. Chem.* **9**, 101–127 (1989).
3. Leaitch, W. R., Hoff, R. M. & McPherson, J. I. *J. atmos. Chem.* **9**, 187–211 (1989).
4. Mickley, R. E., Bottenheim, J. W., Leaitch, W. R. & Evans, W. F. *J. Atmos. Environ.* **23**, 2443–2449 (1989).
5. Oltmans, S. J. *et al. Atmos. Environ.* **23**, 2431–2441 (1989).
6. Sturges, W. T. *et al. Atmos. Environ.* (submitted).
7. Bottenheim, J. W. *et al. J. geophys. Res.* **95**, 18555–18568 (1990).
8. Power, J. F., Sharma, D. K., Langford, C. H., Bonneau, R. & Jousset-Dubien, J. in *Photochemistry of Environmental Aquatic Systems* (ed. Zika, R.) 157–173 (1987).
9. Carey, J. H. & Langford, C. H. *Can. J. Chem.* **53**, 2436–2440 (1975).
10. Behnke, W. & Zetzsch, C. in *Ozone in the Atmosphere* (eds Bojkov, R. D. & Fabian, P.) 519–523 (1989).
11. Zetzsch, C., Pfahler, G. & Behnke, W. *J. Aerosol Sci.* **19**, 1203–1206 (1988).
12. Rancher, J. & Kritz, M. A. *J. geophys. Res.* **85**, 5581–5587 (1980).
13. Finlayson-Pitts, B. J., Livingston, F. E. & Berko, H. N. *Nature* **343**, 622–625 (1990).
14. Patterson, D. E. & Husar, R. B. *Atmos. Environ.* **15**, 1479–1482 (1981).
15. Barrie, L. A. *Atmos. Environ.* **20**, 643–663 (1986).
16. Barrie, L. A. & Barrie, M. J. *J. atmos. Chem.* **11**, 211–226 (1990).
17. Keene, W. C. *et al. Global biogeochem. Cycles* **4**, 407–430 (1990).
18. Duce, R. A., Winchester, J. W. & Van Nahl, T. W. *J. geophys. Res.* **70**, 1775–1799 (1965).
19. Gloersen, P. & Campbell, W. J. *Nature* **352**, 33–36 (1991).
20. Koerner, R. M. in *Polar Bridge, Snow Chemistry Report Appendix E* (Weber, G. R., Dexter, L., Holloway, C. & Buxton, M.) (Key Porter, Toronto, 1990).
21. Yarwood, G., Peng, N. & Niki, H. *J. phys. Chem.* (in the press).
22. Kieser, B. N., Bottenheim, J. W., Niki, H. & Sideris, T. *Atmos. Environ.* (submitted).
23. Berg, W. W., Sperry, P. D., Rahn, A. & Gladney, E. S. *J. geophys. Res.* **88**, 6719–6736 (1983).

ACKNOWLEDGEMENTS. We thank G. Yarwood for discussions, J.C.McC., G.S.H., H.N. and C.H.L. acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada and the Atmospheric Environment Service.

Liquid-hexatic phase transitions in single molecular layers of liquid-crystal films

R. Geer*, T. Stoebe*, C. C. Huang*, R. Pindak†, J. W. Goodby‡, M. Cheng§, J. T. Ho§ & S. W. Hui||

* School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA

† AT&T Bell Laboratories, Murray Hill, New Jersey 07974, USA

‡ School of Chemistry, The University, Hull HU6 7RX, UK

§ Department of Physics and Astronomy, State University of New York at Buffalo, Buffalo, New York 14260, USA

|| Department of Biophysics, Roswell Park Cancer Institute, Buffalo, New York 14263, USA

THEORIES of melting in two dimensions^{1–3} predict that it may have a very different character from the three-dimensional transition. In particular, Halperin and Nelson² proposed that an intermediate hexatic phase, with long-range orientational but not positional order, might intervene between the two-dimensional solid and liquid. Hexatic order has since been identified in three-dimensional liquid-crystal phases⁴. Several liquid-crystal compounds that exhibit a hexatic mesophase can form free-standing films ranging from two to thousands of molecular layers in thickness, permitting experimental investigation of theories of melting in two dimensions, free from the effects of a substrate. We have previously studied transitions between liquid, hexatic and crystal-line phases in films of *n*-heptyl-4'-*n*-pentyloxybiphenyl-4-carboxylate (75OBC), and for films thicker than four layers we observed separate liquid-hexatic transitions for surface and inner layers⁵, revealed by heat-capacity measurements. Here we report that improvements to our measurement technique allow us to identify these transitions in individual molecular layers, even in films just

three layers thick. We also observe a single liquid-hexatic transition in a two-layer film. These films therefore provide model systems for the study of truly two-dimensional phase transitions.

Our experimental technique has been reported elsewhere⁶. Recent improvements have increased the system resolution and stability over wider temperature ranges. Free-standing liquid-crystal films are soap-like films suspended across an aperture (typically 1 cm in diameter)⁷. Unlike in adsorbed systems, there is no substrate to complicate the behaviour of the films near phase transitions. They are created in the smectic-A (SmA) phase which can be viewed as a stack of two-dimensional (2D) liquid layers. Within each layer, the long axes of the rod-like liquid-crystal molecules ($\sim 26 \text{ \AA} \times 6 \text{ \AA}$ in diameter) are, on average, parallel to the layer normal. The average layer spacing is roughly equal to the molecular length. At lower temperatures, the films enter a hexatic-B (HexB) phase in which the molecules acquire three-dimensional (3D) hexatic order (order among local lattice orientations). Below the HexB phase, films enter the crystal-E (CryE) phase which possesses 3D herringbone and crystalline orders^{5,8}.

The free surfaces of free-standing liquid-crystal films induce very unusual behaviour. The surface tension at the film/vapour interface acts to promote order at the surface, in contrast to conventional solids. This, combined with the highly anisotropic nature of smectic phases, often results in the surface layers of the film ordering at a higher temperature than the interior layers⁹⁻¹¹. In Fig. 1a, we show the heat capacity as a function of temperature for an eight-layer 75OBC film. Whereas in ref. 5 we were able to resolve just two transitions in films thicker than four layers, here we find three separate peaks, below which

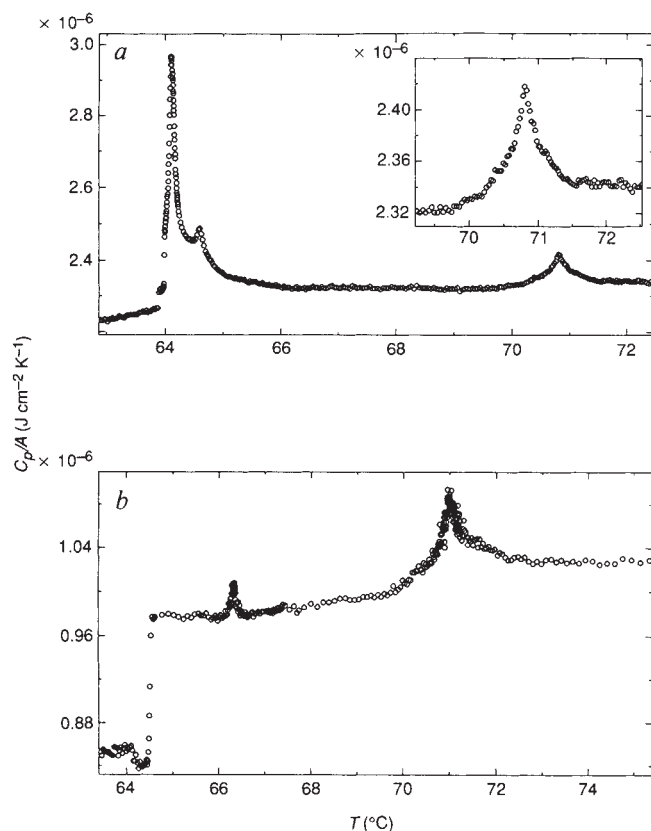


FIG. 1 Heat capacity (C_p) per unit area (A) as a function of temperature for (a) an eight-layer and (b) a three-layer film of 75OBC. In a, the peak at 70.82°C is due to the liquid-hexatic transition in the outermost layers. The peak at 64.60°C results from the liquid-hexatic transition in the layers directly adjacent to the surface layers. The largest peak is due to hexatic ordering in the remaining layers. In b, two peaks result from the surface and interior hexatic ordering, respectively. The abrupt jumps immediately below these peaks correspond to surface hexatic-CryE transitions.

can be seen a pair of abrupt jumps. The high-temperature peak (see inset) results from a SmA-HexB transition of the outermost layers. The second peak (64.8°C) corresponds to the SmA-HexB transition of the layers next to the surface layers. The largest peak, at 64.1°C , is due to hexatic ordering in the remainder of the film. The abrupt jumps indicate a transition from HexB to CryE in the outermost layers⁵. The surface hexatic-CryE transition is first-order with considerable thermal hysteresis⁵. Because of the occurrence of herringbone packing order in the CryE phase, it does not correspond to a simple hexatic-crystal transition, which, according to the two-stage melting theory of Halperin and Nelson, may be continuous.

The layer-by-layer progress of the liquid-hexatic transitions in Fig. 1a results from weak interlayer coupling and the ordering effects of the film/vapour surface tension. The relatively large temperature separation between the surface and next-to-surface liquid-hexatic transitions implies the localization of free surface effects at the outermost layers. In three- and four-layer films, only two liquid-hexatic transitions are present for the surface and interior layers. This is seen in the heat capacity data for a three-layer 75OBC film (Fig. 1b). Two separate liquid-hexatic transitions are evident, one at 71°C and the second at 66.3°C . As in Fig. 1a, the high-temperature peak is associated with the liquid-hexatic transition of the two surface layers. The second peak then corresponds to the liquid-hexatic transition of a single interior layer. The abrupt drop is due to a hexatic-CryE transition of the surface layers, as in Fig. 1a.

We have used electron diffraction¹³, which can probe single hexatic domains, to identify the phases present in these free standing films. For a three-layer film of 75OBC, above 71°C , only a diffuse ring of in-plane scattering is observed, characteristic of a liquid phase. As the film is cooled to a temperature within 2.1°C of the surface hexatic-crystal transition, the ring of scattering remains diffuse but develops sixfold modulation. This is an unambiguous signature of the development of hexatic order. Figure 2 shows a 60° segment of a rotational scan around

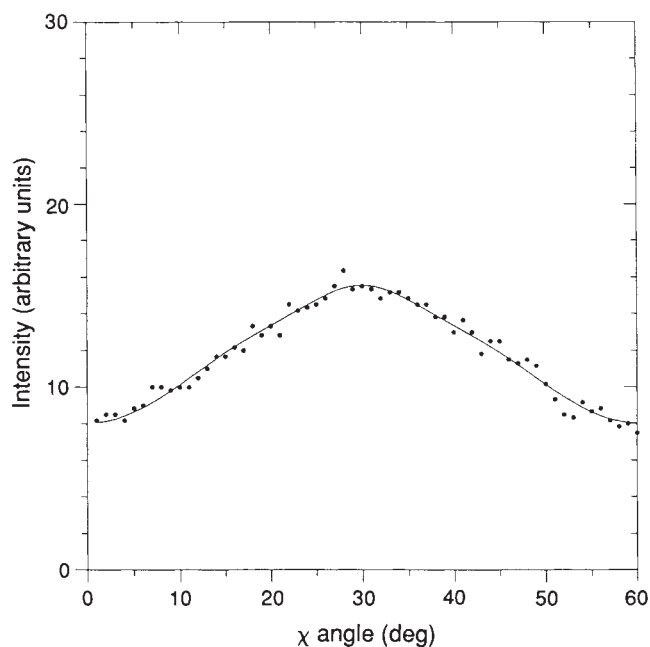


FIG. 2 Diffracted electron intensity for a three-layer 75OBC film. These data were taken 2.06°C above the surface hexatic-CryE transition. This temperature is between the surface-layer and interior-layer transitions. The angle χ refers to a rotation in reciprocal space about the layer normal. For a liquid phase such a scan would produce no change in intensity as a function of χ . The sinusoidal modulation seen here in the diffracted intensity indicates the existence of hexatic order. Such modulation is no longer resolvable above 71°C indicating a transition to a liquid phase. The solid line is a fit similar to that described in ref. 13.

the modulated ring. At this temperature the film is above the lower-temperature heat capacity peak; hexatic order is therefore present in the film up to the higher-temperature heat capacity peak at 71 °C, consistent with the classification of the heat capacity peaks as surface and interior liquid-hexatic transitions⁸. In ref. 5 the peak corresponding to the next-to surface layers was originally identified with the liquid-hexatic transition of the surface layers. Subsequent electron diffraction revealed a small amount of hexatic order above this transition, implying the existence of a higher temperature liquid-hexatic transition. Improvements in the calorimeter have made it possible to resolve the heat capacity peak associated with this transition, as shown in Fig. 1.

The surface liquid-hexatic transitions in Fig. 1a and b represent hexatic ordering on a liquid substrate (the SmA interior layer(s)). In a two-layer film the absence of interior layers results in a single liquid-hexatic transition at 71.1 °C. Figure 3 compares the surface liquid-hexatic transition of a four-layer film with the liquid-hexatic transition of a two-layer film. The latter represents a 2D substrate-free liquid-hexatic transition. Both heat capacity peaks in Fig. 3 are asymmetric, having a higher background heat capacity on the high-temperature side. The peak heights and jumps in background are of comparable magnitude. The two-layer peak, however, is considerably sharper. This may result from the stronger coupling of the layers in the two-layer film than in the surface layers of the four-layer film, which are separated by two liquid-layers. This quality is shared by the

heat capacity peaks associated with the interior liquid-hexatic transitions of Fig. 1a and b, which have narrower full widths at half height than the corresponding surface transitions.

All the measured liquid-hexatic heat capacity peaks are free from thermal hysteresis to within the experimental temperature resolution (± 10 mK). The temperature is measured by a thermometer located at the edge of our film plate. The shape of heat capacity anomalies for the surface liquid-hexatic transitions of Figs 1a, 1b and 3a, and the liquid-hexatic transition in Fig. 3b, are qualitatively similar to, although much sharper than, the theoretical prediction for a 2D continuous liquid-hexatic transition^{12,14}. The theoretical model predicts an increase in the background heat capacity above the transition, caused by the appearance of free topological defects ($k_B/2$ per defect). It also predicts that the transition temperature (temperature at which free defects first appear) lies below the peak position in the heat capacity. Unfortunately, although the electron microscope can be used to detect and quantify the presence of hexatic order, its temperature and instrumental resolution cannot determine the liquid-hexatic transition temperature with sufficient accuracy for comparison to the heat capacity data. Improvements in the resolution of this specialized electron microscope are in progress.

We hope that our heat-capacity measurements of substrate-free, two-layer liquid-crystal films and our observation of the heat-capacity anomaly of the liquid-hexatic transition for a single interior layer of a three-layer film will stimulate further theoretical investigations of 2D melting phenomena for substrate-free systems. □

Received 23 August; accepted 25 October 1991.

1. Kosterlitz, J. M. & Thouless, D. J. *J. Phys.* **C6**, 1181 (1973).
2. Halperin, B. I. & Nelson, D. R. *Phys. Rev. Lett.* **41**, 121 (1978).
3. Young, A. P. *Phys. Rev.* **B19**, 1855 (1977).
4. Pindak, R., Moncton, D. E., Davey, S. D. & Goodby, J. W. *Phys. Rev. Lett.* **46**, 1135 (1981).
5. Geer, R. *et al.* *Phys. Rev. Lett.* **66**, 1322 (1991).
6. Geer, R., Stoebe, T., Pitchford, T. & Huang, C. C. *Rev. scient. Instrum.* **62**, 415 (1991).
7. Young, C. Y., Pindak, R., Clark, N. A. & Meyer, R. B. *Phys. Rev. Lett.* **40**, 773 (1978).
8. Levelut, A. M. *J. Phys. Colloq.* **37**, C3-51 (1976).
9. Pindak, R., Bishop, D. J. & Sprenger, W. O. *Phys. Rev. Lett.* **44**, 1461 (1980).
10. Heinkeamp, S. *et al.* *Phys. Rev. Lett.* **52**, 1017 (1984).
11. Sirota, E. B., Pershan, P. S., Amador, S. & Sorensen, L. B. *Phys. Rev.* **A35**, 2283 (1987).
12. Berker, A. N. & Nelson, D. R. *Phys. Rev.* **B19**, 2488 (1979).
13. Cheng, M., Ho, J. T., Hui, S. W. & Pindak, R. *Phys. Rev. Lett.* **61**, 550 (1988).
14. Solia, S. & Riedel, E. K. *Phys. Rev.* **B23**, 6008 (1981).

ACKNOWLEDGEMENTS. This work was partially supported by the NSF (C.C.H. and J.T.H.), the NIH (S.W.H.) and the Graduate School, University of Minnesota. Fellowship support is gratefully acknowledged from IBM (R.G.), the U.S. Department of Education (T.S.) and the John Simon Guggenheim Memorial Foundation (J.T.H.).

A new class of natural products revealed by 3 β -alkyl steranes in petroleum

Jeremy Dahl, J. Michael Moldowan, Mark A. McCaffrey & Paul A. Lipton

Chevron Oil Field Research Company, PO Box 1627, Richmond, California 94802-0627, USA

THE polycyclic hydrocarbons steranes and hopanes are nearly ubiquitous in petroleum, and their relative resistance to degradation makes them useful as 'biomarkers' for assessing the maturity and history of the oil. Steranes represent the reduced form of sterols, which serve as membrane rigidifiers in eukaryotic organisms; similarly, hopanes derive from hopanols, which serve the same function in prokaryotes. We have identified several homologous series of steranes with alkyl side chains (C_1 to C_6) at the 3 β position. These compounds, when liberated from the polar fractions of the oils by deuterated Raney nickel desulphurization, give fragmentation mass spectra indicating that the 3-alkyl side

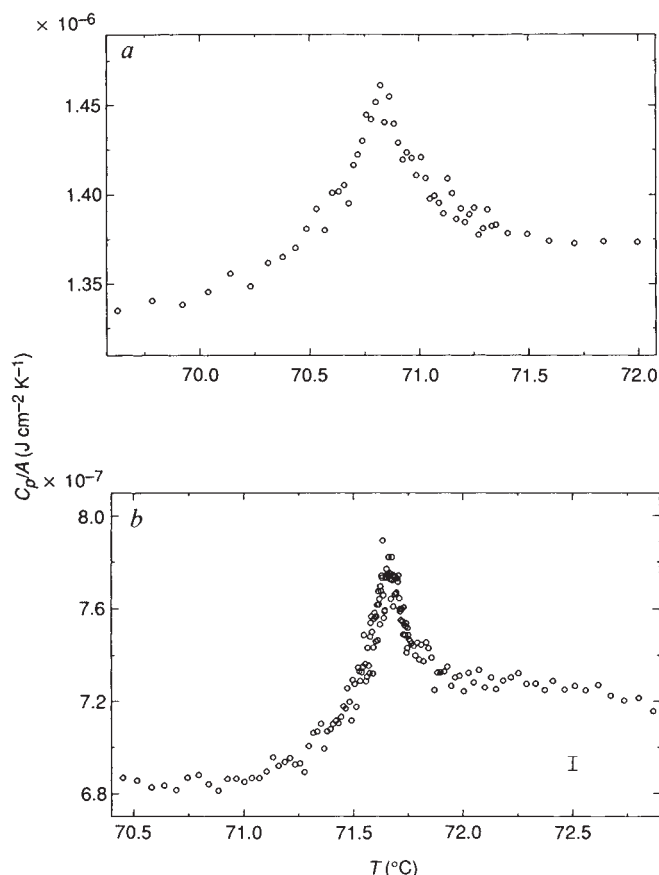


FIG. 3 Heat capacity per unit area as a function of temperature for (a), the surface liquid-hexatic transition in a four-layer 750BC film and (b), the liquid-hexatic transition of a two-layer film. Both peaks are asymmetric with a higher background heat capacity on the high-temperature side. Although for both sets of data hexatic order develops in two layers, the peak is considerably sharper for the two-layer film. (The temperature windows are the same.) This difference is attributable to the interior liquid layers in the four-layer film which may affect the growth of correlations near the transition temperature. The bar in the lower right of b represents a change of $k_B/2$ per particle (see text).

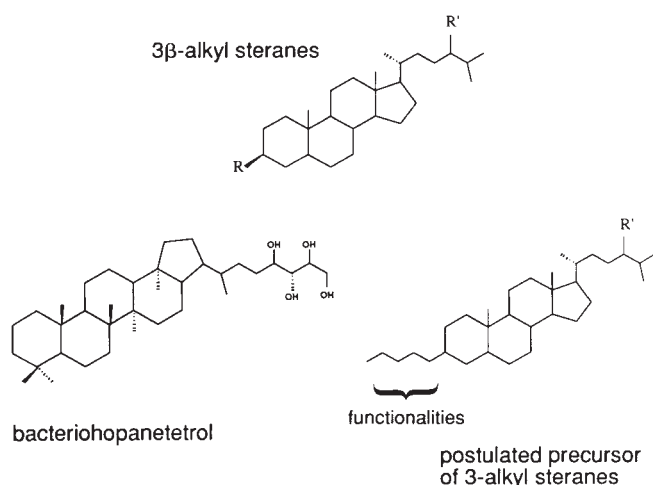


FIG. 1 Structures of 3β-alkyl steranes, bacteriohopanetetrol and hypothetical precursor of 3β-alkyl steranes. Here R is H, CH₃, C₂H₅, ..., C₅H₁₁, ... and R' is H (cholestane), CH₃ (24-methylcholestane), C₂H₅ (24-ethylcholestane).

chains of the precursor steroids contained several functional groups. 3β-pentyl steranes are anomalously abundant in many samples. These data suggest that the precursors represent a new class of steroids, alkylated at the C-3 position with a polyhydroxy *n*-alkane. These precursors may have been formed by the bacterial addition of a ribose sugar to Δ²-sterenes, diagenetic alteration products of steroids synthesized by eukaryotes. 3-alkyl steroids might substitute for hopanols (of prokaryotic origin) in bacterial membranes¹.

The hydrophilic and hydrophobic ends of sterols and the rigid, planar form of these molecules allow them to function as membrane rigidifiers in eukaryotes. With few exceptions, prokaryotes cannot synthesize sterols². Instead, some prokaryotes

synthesize bacteriohopane polyols, which, like sterols, are rigid, planar molecules with hydrophilic and hydrophobic ends suitable for membrane emplacement (Fig. 1)³.

Biomarkers are easily recognized by gas chromatography mass spectrometry (GCMS) from their characteristic fragmentation patterns and isomeric distributions⁴. Analysis by GC-MS-MS tandem mass spectrometry allows detection of very small concentrations of compounds because of the specificity of parent → daughter transitions. Monitoring the parent → daughter transitions shown in Fig. 2 revealed the presence of a series of 3β-alkyl cholestanes (Fig. 3), 3β-alkyl 24-methylcholestanes, and 3β-alkyl 24-ethylcholestanes in the saturate fraction of (1) the Hamilton Dome oil, from the Permian Phosphoria Formation, Wyoming⁵; (2) the 613 oil, from the Miocene Monterey Formation, California⁶; and (3) a source rock extract from the Mowry Shale, Wyoming. Compounds were identified from homology, the presence of characteristic sterane stereoisomer distributions, and co-injection of standard 5α,14α,17α(H) 20*R* and 20*S* and 5α,14β,17β(H) 20*R* and 20*S* stereoisomers of 3β-ethylcholestanes with the sample. The stereoisomers of the standard were prepared by heating authentic 5α,14α,17α(H) 20*R* 3β-ethylcholestane provided by Summons⁷ in a sealed tube with 5% palladium on activated carbon at 260 °C for 24 h (ref. 8).

The 3β-alkyl group extends to at least hexyl, and there are indications that higher homologues are also present. With the exception of 3β-pentyl steranes, the concentrations of the 3β-alkyl steranes decrease steadily with increasing side-chain length. This indicates that the side chain is a normal, rather than a branched alkyl group⁹. The relative amounts of cholestanes, 24-methylcholestanes and 24-ethylcholestanes are approximately the same for each 3β-alkyl sterane homologue and for steranes with no A-ring alkylation, as are the distributions of stereoisomers. This suggests a common precursor for the 3β-alkyl and 3-non-alkylated steranes. 3β-methyl- and 3β-ethylcholestanes have been reported by Summons and Capon^{7,10}, and they suggested that the similarity of isomeric

3-alkyl cholestanes			3-alkyl 24-methylcholestanes		
R'	R	Transition	R'	R	Transition
Fragment A					
H	CH ₃	386 → 231	CH ₃	CH ₃	400 → 231
H	C ₂ H ₅	400 → 245	CH ₃	C ₂ H ₅	414 → 245
H	C ₃ H ₇	414 → 259	CH ₃	C ₃ H ₇	428 → 259
H	C ₄ H ₉	428 → 273	CH ₃	C ₄ H ₉	442 → 273
H	C ₅ H ₁₁	442 → 287	CH ₃	C ₅ H ₁₁	456 → 287
Fragment B					
H	CH ₃	386 → 262	CH ₃	CH ₃	400 → 276
H	C ₂ H ₅	400 → 262	CH ₃	C ₂ H ₅	414 → 276
H	C ₃ H ₇	414 → 262	CH ₃	C ₃ H ₇	428 → 276
H	C ₄ H ₉	428 → 262	CH ₃	C ₄ H ₉	442 → 276
H	C ₅ H ₁₁	442 → 262	CH ₃	C ₅ H ₁₁	456 → 276

3-alkyl 24-ethylcholestanes		
R'	R	Transition
Fragment A		
C ₂ H ₅	CH ₃	414 → 231
C ₂ H ₅	C ₂ H ₅	428 → 245
C ₂ H ₅	C ₃ H ₇	442 → 259
C ₂ H ₅	C ₄ H ₉	456 → 273
C ₂ H ₅	C ₅ H ₁₁	456 → 287
Fragment B		
C ₂ H ₅	CH ₃	414 → 290
C ₂ H ₅	C ₂ H ₅	428 → 290
C ₂ H ₅	C ₃ H ₇	442 → 290
C ₂ H ₅	C ₄ H ₉	456 → 290
C ₂ H ₅	C ₅ H ₁₁	470 → 290

Fragment A

Fragment B

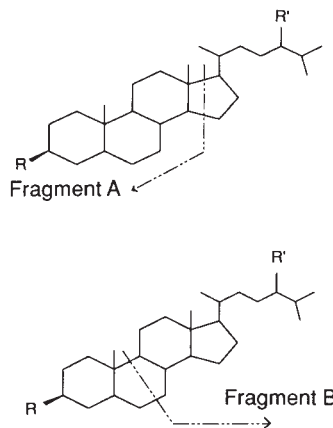
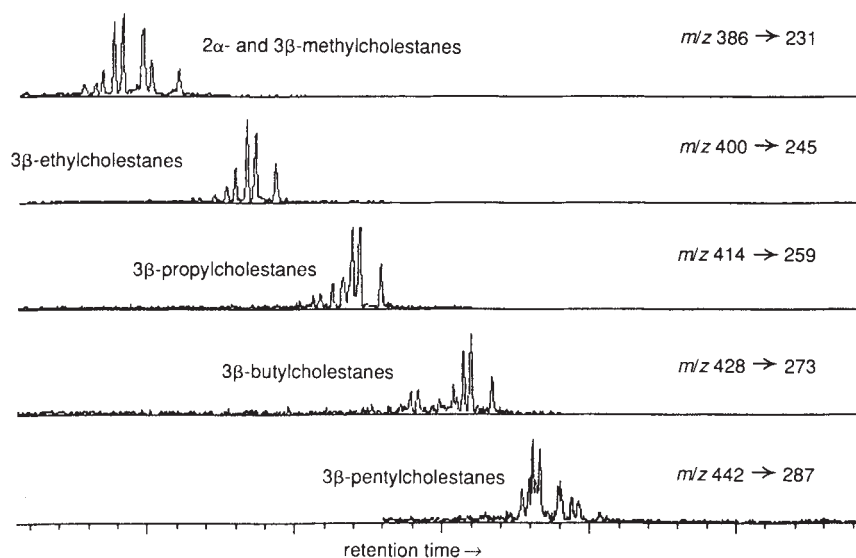


FIG. 2 Parent-to-daughter transitions of the saturate fraction of oils and rock extracts, separated by high-pressure liquid chromatography, were monitored by GC-MS-MS (Finnigan TSQ70) for 3β-alkyl steranes (transitions listed in table). Fragment A parent-to-daughter transitions show increasing alkyl side-chain length at the 3 position, whereas fragment B shows alkylation at the 24 position. Experimental conditions: GC carrier gas was H₂ at a flow velocity of 38 cm s⁻¹ through a J&W fused silica capillary, DB-1 (methyl silicone); 0.25 μm film thickness, 0.25 mm internal diameter and 60 m length. The temperature program was isothermal for 2 min at 100 °C, then ramped at 15 °C min⁻¹ to 150 °C, 2 °C min⁻¹ to 320 °C and isothermal for 20 min at 320 °C with injector and transfer line temperatures of 320 °C. Ionization was by electron impact at 70 eV in the ion source and by collision-induced decomposition at 15 eV using argon at 0.5–0.6 mtorr in the second quadrupole.

FIG. 3 Homologous series of cholestane isomers with alkylation at the 3 position are detected in mass chromatograms constructed from GC-MS-MS parent-to-daughter transitions of the saturate fraction of the Hamilton Dome oil from Wyoming. 3β -ethylcholestanes were confirmed by coinjection of authentic standards. Transitions from parent to ion at mass/charge ratio $m/z=262$, where m is in a.m.u. and z in e (see Fig. 2) confirmed alkylation in the A-ring for all cholestanes. Similar homologous series are observed for 3-alkyl 24-methyl and 3-alkyl 24-ethylcholestanes. Although coelution experiments using 3β -ethylcholestane show the 3β -alkyl steranes to have 3β stereochemistry, one cannot assume from this that the stereochemistry of the polyfunctional 3-alkyl side chain is 3β in the natural product precursor. This is because stereoisomerization during diagenesis is common. If, however, these precursors are functional bacterial membrane components, a 3β (equatorial) side chain, which extends the linear sense of the steroid carbon skeleton, would be favoured over a 3α (axial) one.



distribution between the 3β -alkyl and non-alkylated steranes was the result of one or two bacterial methylations of Δ^2 -sterenes, diagenetic intermediates in the formation of steranes from sterols^{7,10}. In addition, they reported 2α -methyl steranes, supporting the contention of a Δ^2 -sterene precursor.

Sulphur-bound 3β -alkyl steranes provide other clues to the biological precursors of these compounds. Organo-sulphur compounds are formed during early diagenesis by the reaction of sulphur, polysulphides or H_2S with organic compound functionalities¹¹⁻¹⁴. Raney nickel desulphurization¹⁵⁻¹⁷ can be used to break carbon-sulphur bonds and liberate hydrocarbons from their sulphur-bound counterparts. 3β -alkyl steranes were generated by Raney nickel desulphurization of the polar maltene fraction of the 613 Monterey oil. Use of deuterium-labelled Raney nickel resulted in 3β -alkyl cholestanes with one to five

deuteriums in the A fragments (Fig. 2) and a low abundance of only slightly deuterated (D_0 - D_2) B fragments, an observation suggesting multiple sulphur bonding on the 3β -alkyl side chain. The results of the Raney nickel and deuterated Raney nickel desulphurizations indicate that the 3-alkyl steranes existed before, or during, early diagenesis and that their precursors had functional groups mainly in the 3-alkyl side chain. Furthermore, the 3β -alkylcholestanes liberated by the desulphurization show a strong predominance of 3β -pentylcholestane over the lower homologues (Fig. 4). This pattern implies that there is a highly functionalized precursor for 3β -pentylcholestane which preferentially reacts with sulphur during diagenesis.

To a lesser extent, the 3β -pentylcholestanes are also anomalously abundant in the original saturates in comparison with the trend of decreasing concentration shown by the other homologues of the Monterey oil. The relative abundance of moieties with a C_5 side chain is analogous to the homologous series of $17\alpha,21\beta(H)$ homohopanes in this oil (Fig. 4) and other thermally immature oils derived from anoxic, marine source rocks¹⁸. Anomalous abundances of pentakishomohopanes (C_{35}), relative to other homohopanes (C_{31} - C_{34}) in saturate fractions, result from the selective preservation of the carbon skeleton of the C_{35} precursor to all the homohopanes¹⁸, bacteriohopanetetrol and related polyols¹⁹. The predominance of C_{35} homohopane in desulphurized fractions is generally believed to result from the selective incorporation of bacteriohopane polyols into the sulphur-bound fraction^{20,13}. Bacteriohopanetetrol is synthesized by bacterial addition of a C_5 sugar (a D-ribose derivative) to a C_{30} hopanoid (possibly hop-22(29)-ene)¹, and therefore has a tetrahydroxy side-chain with ample sites for reaction with reduced sulphur species. By analogy, we suggest that the precursors to 3β -alkyl steranes are synthesized by bacterially mediated addition of a C_5 sugar to a

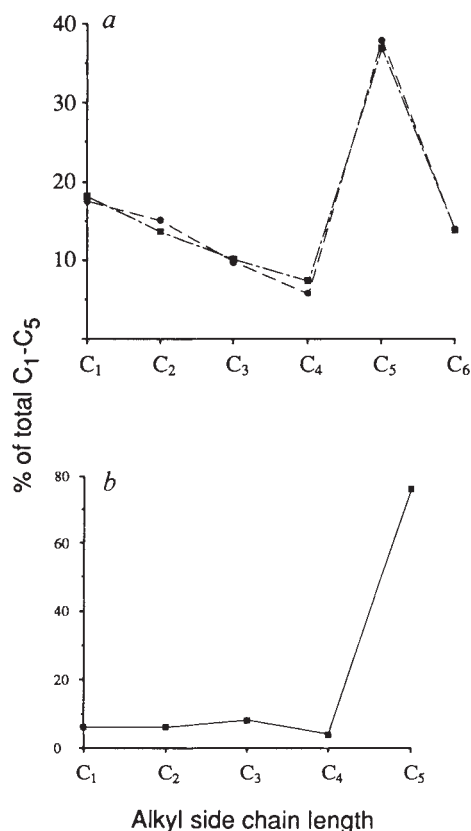


FIG. 4 a, The distribution of 3β -alkylcholestanes in the saturate fraction generated by Raney nickel desulphurization of the polar maltene fraction of the Monterey 613 oil. Distribution was determined by (1) the total number of counts on the total ion chromatogram for each transition and (2) the peak area of the $5\alpha,14\alpha,17\alpha(H)$ 20R isomer of each homologue in the mass spectrum. Note that 3β -pentylcholestanes are the most abundant 3β -alkyl sterane homologues. b, The distribution of homohopanes in the saturate fraction generated by Raney nickel desulphurization of the polar maltene fraction of the Monterey 613 oil. These data suggest that the 3-alkyl steranes (like the bacteriohopanes) are formed by the addition of a C_5 sugar; in the case of 3-alkyl steranes, the addition is probably to a Δ^2 -sterene. A slightly elevated C_6 homologue in the Raney nickel product is also noteworthy because it may indicate some 3-alkylations of steranes with a C_6 sugar.

pre-existing steroid, possibly a Δ^2 -sterene. The polyfunctional steroid precursors thus formed may function as surrogates for the bacteriohopanols in bacterial membranes. This would be analogous to known examples of incorporation of sterols into the membranes of organisms that cannot biosynthesize them. For instance, the protozoan *Tetrahymena*, which synthesizes tetrahymanol, will substitute sterols if they are available²¹. If our postulated origin of 3 β -alkyl steranes is correct, the membranes of certain bacteria living in environments rich in sterol degradation products should contain 3-alkyl steroids polyfunctionalized on the 3-alkyl group. Should 3 β -alkyl steranes be shown to be geochemical products of steroid alkylation by bacteria for incorporation into their membranes, we suggest the trivial name 'bacteriosterane' to describe their carbon skeleton, in analogy to the term 'bacteriohopane'. Finally, 3 β -carboxysteranes, which are abundant in some sediments²², may be formed by oxidative cleavage of the polyfunctional chain of 3-alkyl steroids, and thus are linked by a common precursor to the 3 β -alkyl steranes. $\square$

Received 27 August; accepted 15 October 1991.

1. Flesch, G. & Rohmer, M. *Eur. J. Biochem.* **175**, 405–411 (1988).
2. Rohmer, M., Bouvier-Nave, P. & Ourisson, G. *J. gen. Microbiol.* **130**, 1137–1150 (1984).
3. Ourisson, G., Albrecht, P. & Rohmer, M. *Pure appl. Chem.* **51**, 709–729 (1979).
4. Philp, R. P. *Fossil Fuel Biomarkers, Applications and Spectra*, 1–294 (Elsevier, Amsterdam, 1985).
5. Seifert, W. K. & Moldowan, J. M. *Geochim. cosmochim. Acta* **45**, 783–794 (1981).
6. Peters, K. E. & Moldowan, J. M. *Org. Geochem.* **17**, 47–61 (1991).
7. Summons, R. E. & Capon, R. J. *Geochim. cosmochim. Acta* **55**, 2391–2395 (1991).
8. Moldowan, J. M. *et al. Science* **247**, 309–312 (1990).
9. Moldowan, J. M., Seifert, W. K. & Gallegos, E. G. *Geochim. cosmochim. Acta* **47**, 1531–1534 (1983).
10. Summons, R. E. & Capon, R. J. *Geochim. cosmochim. Acta* **52**, 2733–2736 (1988).
11. Brassell, S. C., Eglinton, G. & Mo, F. J. *Org. Geochem.* **10**, 927–941 (1986).
12. Sinninghe Damsté, J. S., Rijpstra, W. I. C., Kock-van Dalen, A. C., de Leeuw, J. W. & Schenck, P. A. *Geochim. cosmochim. Acta* **53**, 1343–1355 (1989).
13. de Leeuw, J. W. & Sinninghe Damsté, J. S. in *Geochemistry of Sulfur in Fossil Fuels* (eds Orr, W. L. & White, C. M.) 417–443 (American Chemical Society, Washington, DC, 1990).
14. Fukushima, K., Yasukawa, M., Muto, N., Uemura, H. & Ishiwatari, R. *Org. Geochem.* **17** (in the press).
15. Schmid, J.-C. thesis, Univ. Louis Pasteur de Strasbourg (1986).
16. Sinninghe Damsté, J. S. *et al. Geochim. cosmochim. Acta* **51**, 2369–2391 (1987).
17. Trillief, S. thesis, Univ. Louis Pasteur de Strasbourg (1987).
18. Moldowan, J. M. *et al. in Biological Markers in Sediments and Petroleum* (eds Moldowan, J. M., Albrecht, P. & Philp, R. P.) (Prentice Hall, Englewood Cliffs, New Jersey, in the press).
19. Rohmer, M., Bissleret, P. & Neunlist, S. in *Biological Markers in Sediments and Petroleum* (eds Moldowan, J. M., Albrecht, P. & Philp, R. P.) (Prentice Hall, Englewood Cliffs, New Jersey, in the press).
20. Sinninghe Damsté, J. S., Rijpstra, W. I. C., de Leeuw, J. W. & Schenck, P. A. in *Advances in Organic Geochemistry 1987, Org. Geochem.* (eds Novelli, L. & Mattavelli, L.) 593–606 (1988).
21. Conner, R. L., Landrey, J. R., Burns, C. H. & Mallory, F. B. *J. Protozool.* **15**, 600–605 (1968).
22. Dany, F., Riolo, J., Trendel, J.-M. & Albrecht, P. *J. chem. Soc. Chem. Commun.* 1228–1230 (1990).

ACKNOWLEDGEMENTS. We thank P. Sundararaman for discussion and R. E. Summons for standards.

Control of dissolution rates of orthosilicate minerals by divalent metal–oxygen bonds

William H. Casey* & Henry R. Westrich†

* Land, Air and Water Resources, University of California, Davis, California 95616, USA

† Geochemistry Research Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

COMMON rock- and soil-forming minerals are complicated structures of varying composition. Despite some encouraging progress^{1,2} there is as yet no comprehensive rationale for predicting the dissolution rates of these minerals. Here we test the hypothesis³ that dissolution rates of compositionally distinct orthosilicate minerals scale in a fashion similar to rates of water exchange around the corresponding dissolved, divalent cation. Although dissolution rates span several orders of magnitude, the hypothesis is sustained. Minerals containing alkaline-earth cations dissolve at rates that correlate with ionic size, whereas minerals containing first-row transition metals dissolve at rates that vary with the number of cation *d*-electrons. Both types of behaviour are consistent with the control of dissolution rate by the character of the

bonds between the divalent cation and neighbouring oxygen atoms. This result supports the proposed link^{3–6} between the mechanisms of mineral dissolution and the mechanisms by which a dissolved metal exchanges ligands. With this link it may be possible to predict dissolution rates for other nearly isostructural minerals that vary in composition.

Several authors^{3–6} have suggested that there are similarities between the mechanisms by which a cation leaves a dissolving mineral surface and those by which a dissolved complex exchanges ligands. In proton-promoted dissolution, bridging oxygens at the mineral surface are replaced with water molecules or hydroxyl ions. This replacement can follow rapid protonation steps⁶, but ultimately a hydrated ion or hydroxy-complex is released from the mineral to the aqueous solution. Likewise, exchange of the hydration waters around a dissolved metal requires the destruction and reformation of metal–oxygen bonds.

The coordination chemistry of metals in solution and near the mineral surface is often similar. Magnesium ion, for example, is coordinated to six oxygens both in forsterite (Mg_2SiO_4) and in the hydrated ion ($\text{Mg}(\text{H}_2\text{O})_6^{2+}$). The Mg–O bond lengths in forsterite (2.101–2.126 Å)⁷ are similar to that in the hydrated ion (2.10 Å)⁸. Of common orthosilicates considered here, only willemite (Zn_2SiO_4) has a coordination chemistry that is considerably different from the hydrated ion³.

Unlike rates of mineral dissolution, however, reactivity trends for ligand-exchange rates are well defined⁸. Rates of solvent exchange correlate with ion size for hydrated alkaline-earth ions ($\text{Ca}(\text{H}_2\text{O})_6^{2+} > \text{Mg}(\text{H}_2\text{O})_6^{2+} > \text{Be}(\text{H}_2\text{O})_4^{2+}$). Ion size is important because large cations more easily accommodate the extra ligand in associative exchange than small cations. Conversely, rates vary with the number of *d*-electrons for hydrated first-row transition metals, where bonds to oxygens have strong directional preference due to unequal filling of oriented bonding and antibonding orbitals by *d*-electrons. For these metals, exchange rates correlate with the ligand-field stabilization energies^{8,9}; that is, $\text{Zn}(\text{H}_2\text{O})_6^{2+} \approx \text{Mn}(\text{H}_2\text{O})_6^{2+} > \text{Fe}(\text{H}_2\text{O})_6^{2+} > \text{Co}(\text{H}_2\text{O})_6^{2+} > \text{Ni}(\text{H}_2\text{O})_6^{2+}$. The metals that are most stable in octahedral coordination to oxygens resist the distortion that accompanies replacement of one coordinated ligand with another^{5,9}. Divalent manganese and zinc have no excess of bonding over antibonding orbitals (the electronic state of Mn^{2+} is d^5 , that of Zn^{2+} is d^{10}) and are thus not particularly stable in octahedral coordination to oxygens. These metals react at similar, relatively rapid, rates.

One expects a similar reactivity trend in silicate mineral weathering if the bonds between the divalent metal and oxygens exert considerable control over the reaction, and if the mechanisms of dissolution resemble mechanisms of ligand exchange. Orthosilicate minerals are ideal to test this idea because: (1) divalent substitutions do not markedly affect bond lengths or angles in the silicate anion¹⁰; (2) the minerals can be acquired as pure compounds, so that energetic differences among cation sites can be ignored²; and (3) the minerals are similarly affected by slight variations in solution pH^{11–14}. For minerals of the olivine and willemite classes, the relative order of dissolution rates is predicted to be: $\gamma\text{-Ca}_2\text{SiO}_4 > \text{Mg}_2\text{SiO}_4$ (forsterite) $> \text{Be}_2\text{SiO}_4$ (phenakite) for alkaline-earth orthosilicates and Zn_2SiO_4 (willemite) $\approx \text{Mn}_2\text{SiO}_4$ (tephroite) $> \text{Fe}_2\text{SiO}_4$ (fayalite) $> \text{Co}_2\text{SiO}_4 > \text{Ni}_2\text{SiO}_4$ (liebenbergite) for orthosilicates with first-row transition metals.

To test the hypothesis, dissolution rates for the minerals were directly measured or compiled from published results^{11,12,14}. Most compounds were synthesized (Table 1) to ensure chemical and mineralogic purity. Products of these syntheses were ground under ethanol, sieved to isolate the size fraction from 25 to 75 μm and ultrasonically agitated in ethanol to remove ultrafine particles. The surface of each powder was measured using N_2/He gas mixtures and the BET adsorption isotherm. We confirmed the mineral purity by X-ray diffraction analysis and by microscopic examination of the resulting powder. Minor

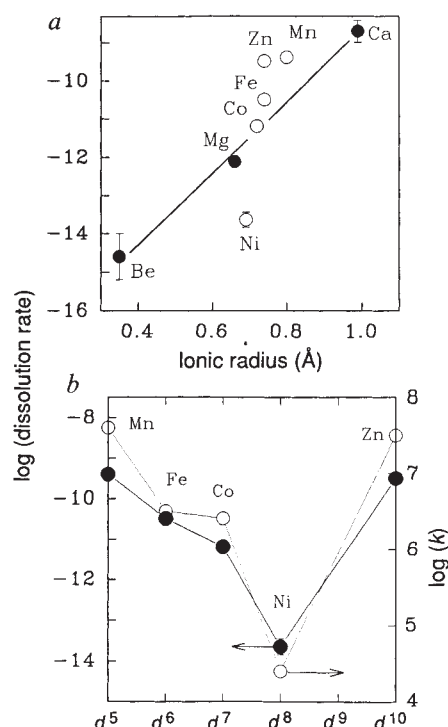


FIG. 1 Dissolution rates of orthosilicate minerals (identified by the dominant divalent cation) at pH $\approx$ 2 and a temperature of 25 °C. *a*, Rates for minerals containing alkaline-earth cations (filled symbols) vary with ionic radius of the divalent cation. Rates for minerals containing first-row transition metals (open symbols) vary considerably, although there is little variation in ionic radius. *b*, Dissolution rates of orthosilicate minerals containing transition metals (filled symbols, left-hand axis) compare well with the first-order rate coefficient¹⁵, k (s^{-1}), describing water exchange from the solvent into the hydration sphere of the corresponding metal ion (open symbols, right-hand axis). Measured uncertainties are commonly within the size of the plotted symbol.

oxide impurities in fayalite and liebenbergite were removed by treatments with high-density liquids and by repeated magnetic separation. The final powders contained less than 1–2% by volume of impurities, most of which were sealed in silicate crystals and inaccessible to the aqueous solution.

We measured the dissolution rates in chloride solutions using the pH-stat method¹¹ over the range $2 < \text{pH} < 5$, or by using a chemostat cell^{12,14}. The accuracy and precision of any single rate estimate by means of the pH-stat method was generally much better than $\pm 10\%$. The accuracy was determined by analysing samples of the solution for dissolved metal concentration as a function of time. These data were linearly regressed to estimate a batch dissolution rate which is compared with the rate from the pH-stat method. The accuracy was worse for cases

TABLE 1 Synthesis conditions

Mineral	Starting material	Temperature (°C)	Atmosphere	Reference
Ca-olivine ($\gamma\text{-Ca}_2\text{SiO}_4$)	CaO-SiO ₂ gel	1,575–1,500	Air	this study
Forsterite (Mg_2SiO_4)	natural	—	—	12, 14
Phenakite (Be_2SiO_4)	natural	—	—	this study
Willemite (Zn_2SiO_4)	ZnO, SiO ₂	1,580–1,530	Air	11
Tephroite (Mn_2SiO_4)	MnCO ₃ , SiO ₂	1,375–1,325	N ₂	this study
Fayalite (Fe_2SiO_4)	Fe-oxalate, SiO ₂	1,275–1,175	CO/CO ₂	this study
Co-olivine (Co_2SiO_4)	CoO, SiO ₂	1,450–1,400	Air	this study
Liebenbergite (Ni_2SiO_4)	NiO, SiO ₂ , Mo-flux	1,500–1,100	Air	this study

Most minerals were crystallized slowly from a melt and all syntheses were conducted in platinum crucibles. Ni_2SiO_4 and $\gamma\text{-Ca}_2\text{SiO}_4$ were synthesized by subsolidus reaction. Lithium molybdate was used as a flux to crystallize Ni_2SiO_4 . The crystalline Ni_2SiO_4 was separated from this flux by magnetic separation. Fayalite was synthesized at an oxygen fugacity of $\sim 10^{-9.5}$ atm. Natural and synthetic samples of willemite were compared in ref. 11.

such as $\gamma\text{-Ca}_2\text{SiO}_4$, where the concentration varies nonlinearly in a short period of time as the reaction proceeds to completion. This mineral also dissolves so rapidly that the solution must be vigorously agitated to avoid control of the reaction by solute diffusion and not by the kinetics of reaction on the mineral surface. We thus estimate an overall uncertainty of a factor of two for $\gamma\text{-Ca}_2\text{SiO}_4$.

A summary of experimental conditions is presented in Table 2. All experiments, except those on phenakite, willemite¹¹ and forsterite^{12,14}, were conducted under a nitrogen atmosphere to maintain dissolved O₂ concentrations less than 0.2 p.p.m. as measured with an oxygen electrode. The oxidation state of cations in phenakite, willemite and forsterite are not affected by oxygen in water. Because such small fractions of the liebenbergite and phenakite were dissolved, additional long-term experiments were conducted in chemostat cells to monitor the variation in rates with time, and this variation is used as an overall uncertainty in Table 2.

As predicted (Fig. 1*a*), rates vary with ion size for minerals containing alkaline-earth cations in the relative order $\text{Ca}_2\text{SiO}_4 > \text{Mg}_2\text{SiO}_4 > \text{Be}_2\text{SiO}_4$. For minerals containing transition metal cations, dissolution rates vary considerably, although there is very little change in ionic radius (Fig. 1*a*). The order of reactivity $\text{Zn}_2\text{SiO}_4 \approx \text{Mn}_2\text{SiO}_4 > \text{Fe}_2\text{SiO}_4 > \text{Co}_2\text{SiO}_4 > \text{Ni}_2\text{SiO}_4$ reflects the ligand-stabilization energies of the divalent cation. Tephroite and willemite dissolve at similar rapid rates, whereas liebenbergite dissolves many orders of magnitude more slowly (Fig. 1*b*).

Except for phenakite, all minerals dissolve to release metals in stoichiometric proportions, as has been found previously in

TABLE 2 Experimental conditions

Mineral	Duration (s)	pH	Mass consumed	Rate ($\pm 1\sigma$) (mol cm ⁻² s ⁻¹)	d(log rate)/dpH ($\pm 1\sigma$)
$\gamma\text{-Ca}_2\text{SiO}_4$	500	2.0	50%	$2.0(\pm 1.0) \times 10^{-9}$	—
Mg_2SiO_4	—	1.95–2.0	—	$7.5(\pm 1.5) \times 10^{-13}$	(−0.54) (refs 12, 14)
$\text{Be}_2\text{SiO}_4^*$	1.1×10^6	2.0	0.04%	$2.5(\pm 2.0) \times 10^{-15*}$	(−0.3 $\pm$ 0.05)†
Zn_2SiO_4	—	1.90	<(70–90)%	$3.2(\pm 0.1) \times 10^{-10}$	(−0.4 to 0.5) (ref. 11)
Mn_2SiO_4	7,000	2.0	0.16%	$4.0(\pm 0.5) \times 10^{-10}$	−0.4 $\pm$ 0.04
Fe_2SiO_4	26,000	2.0	14%	$3.0(\pm 0.8) \times 10^{-11}$	−0.15 to −0.4
Co_2SiO_4	350,000	2.0	5%	$6.1(\pm 0.8) \times 10^{-12}$	−0.5 $\pm$ 0.06
$\text{Ni}_2\text{SiO}_4^\ddagger$	17,000	2.0	0.01%	$2.3(\pm 1) \times 10^{-14}$	(−0.43 $\pm$ 0.04)‡

* Uncertainty estimated as the difference in the total range of rates observed over 1.1×10^6 s.

† Uncertainty estimated as the difference between rates measured by pH-stat method (17,000 s) and chemostat experiments conducted at 50 °C for 4×10^6 s. An activation energy of 10 kcal mol⁻¹ (refs 11, 13) was used in converting these rates to 25 °C.

‡ Determined at 50 °C.

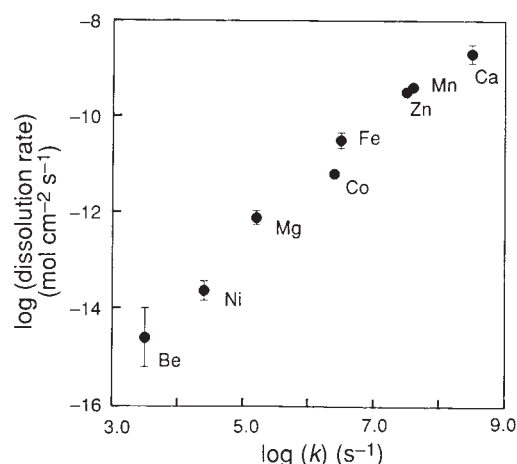


FIG. 2 Dissolution rates for endmember orthosilicate minerals at pH=2 and 25 °C, plotted against the first-order rate coefficient, k (s^{-1}), for water exchange from the solvent into the hydration sphere of the corresponding dissolved cation. The mineral and ions are identified by the divalent cation (for example, Mn refers to Mn_2SiO_4 and $\text{Mn}(\text{H}_2\text{O})_6^{2+}$).

acid conditions^{12–14}. Phenakite dissolves to release silicon in slight excess of that predicted from the mineral stoichiometry. We attribute this to the dissolution of small amounts of feldspar inclusions in the natural mineral sample. Differences in the pH-dependence of dissolution rates among the minerals are relatively small (Table 2) and do not account for the differences shown in Fig. 2.

The wide range of rates and the striking similarity between the reactivities for mineral dissolution and solvent exchange (Fig. 2) indicate that dissolution proceeds along similar pathways to solvent exchange around the divalent cation. In these experimental conditions, the character of the divalent cation in a silicate structure influences weathering rates considerably. The extent to which similar relations will be observed in high-pH solutions, where silicon-oxygen bonding may be more important¹ to the hydrolysis kinetics and different surface reactions predominate¹⁴, remains uncertain.

The importance of our result extends beyond the orthosilicate mineral class. A similar order of reactivities is expected for other silicate materials that differ primarily in the divalent cation, that have a coordination chemistry similar to the hydrated cation, and for which dissolution rates are limited by surface reactions and not by solute transport. We predict that pyroxene minerals of varying composition, or even silicate glasses of similar structure but varying composition, will show a similar order of reactivities. □

Locomotion by jumping in the Mediterranean fruit-fly larva *Ceratitis capitata*

David P. Maitland

Department of Physiology, Medical School,
University of the Witwatersrand, Johannesburg 2193, South Africa

IN contrast to terrestrial locomotion in animals with rigid internal or external skeletons (such as insects or vertebrates), caterpillar crawling is slow and energetically costly¹. Biomechanical constraints may be responsible for the high cost of crawling locomotion¹, but the caterpillar body plan, which uses hydraulic-based skeletal and locomotory systems, does not necessarily preclude the possibility of high-speed locomotion as has been suggested¹. I report here how the last-instar larva of the Mediterranean fruit-fly jumps. Jumping increases the maggot's speed to 0.5 m s^{-1} per jump, or by 200-fold over crawling. Jumping occurs at a stage when the maggot is particularly vulnerable to parasitization and predation. The fly larva provides the only known example of jumping by a soft-bodied legless organism.

When ready to pupate, Mediterranean fruit-fly larvae (Fig. 1) bore out of their host fruit to emerge onto its surface or drop to the ground. Here they prepare to jump (Fig. 2). Jump preparation can be divided into two distinct but continuous phases: loop formation and tension generation.

In the first phase, the maggot arches itself by contraction of ventral longitudinal muscles, bringing the head in contact with the tail to allow a pair of mouth hooks to grip the cuticle below the posterior spiracles. The maggot is now standing in an unusually tight but unknicked loop (Fig. 2a).

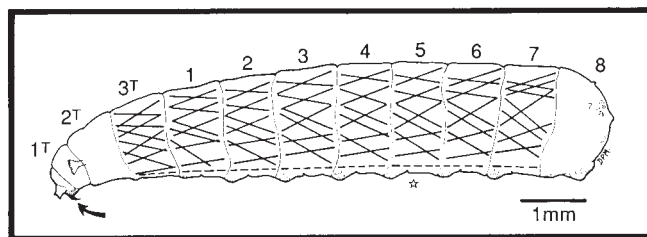


FIG. 1 Last-instar larva of the Mediterranean fruit-fly *Ceratitis capitata*, weight 17 mg (± 1.9 s.d.; $n=20$) and length 8.5 mm (± 0.25 ; $n=10$). After the head, with protruding mandibular sclerites¹² or 'mouth hooks' (curved arrow), there are 11 visible body segments, namely 3 thoracic segments (1T–3T) and 8 abdominal segments (1–8). Abdominal segment 5 is marked with a star, see Fig. 2 legend. The maggot's skin is 25 μm thick and is kept taut by a standing internal hydrostatic pressure of 22 mm Hg (± 4.3 ; $n=12$) (indirect measure obtained by re-inflating decapitated maggots to their original dimensions). The body wall is fibre-reinforced with a webbing of crossed-helical non-muscular fibre bundles 51 μm (± 8 ; $n=10$) in diameter arranged at an angle of 75° (± 2.3 ; $n=11$) to the horizontal. Anchored to the skin are rows of segmental muscle bands (those on the maggot's left are shown; solid diagonal lines). Five rows are inclined to the left and 5 rows are inclined to the right at angles of between 3° and 40° to the long axis ($\bar{x}=19.5 \pm 10.4^\circ$; $n=70$). Thus, the muscle bands collectively describe right- and left-handed helices around the maggot's body. The first row of ventral longitudinal muscles is shown (dotted line). The muscles constitute 16% (or 2.7 ± 0.3 mg; $n=7$) of the maggot's body mass, and the skin represents 13% (or 2.15 ± 0.2 mg; $n=7$). The body-wall stress-strain relationship for experimentally inflated maggot skins (methods modified after ref. 13) revealed that the energy required for elastic deformation of the maggot skin could be almost completely recovered (the elastic limit¹⁴) up to a maximum circumferential strain of 29% (internal hydrostatic pressure of 82 ± 9 mm Hg; $n=9$) (circumferential stress of 0.4 ± 0.06 MPa $n=9$). The potential 'strain energy storage'¹⁴ capacity for maggot skin is therefore about 31 J kg⁻¹.

Received 5 August; accepted 25 October 1991.

- Brady, P. V. & Walther, J. V. *Geochim. cosmochim. Acta* **53**, 2823–2830 (1989).
- Schott, J., Berner, R. A. & Sjöberg, E. L. *Geochim. cosmochim. Acta* **45**, 2123–2135 (1981).
- Casey, W. H. J. *Colloid Interface Sci.* **135**, (in the press).
- Stumm, W. & Wollast, R. *Rev. Geophys.* **28**, 53–69 (1990).
- Burns, R. G. *Mineralogical Applications of Crystal Field Theory* (Cambridge University Press, London, 1970).
- Casey, W. H., Lasaga, A. C. & Gibbs, G. V. *Geochim. cosmochim. Acta* **54**, 3369–3378 (1990).
- Bish, D. L. & Burnham, C. W. *Am. Mineral.* **69**, 1102–1109 (1984).
- Burgess, J. *Ions in Solution* (Wiley, New York, 1988).
- Wilkins, R. G. *The Study of Kinetics and Mechanisms of Reactions of Transition Metal Complexes* (Allyn-Bacon, Boston, 1974).
- Brown, G. E. in *Orthosilicates* (ed. Ribbe, P.) 175–365 (Mineralogical Society of America, Washington DC, 1983).
- Terry, B. & Monhemius, A. J. *Metall. Trans.* **B14**, 335–326 (1983).
- Blum, A. & Lasaga, A. C. *Nature* **331**, 431–433 (1988).
- Grandstaff, D. E. *Proc. 3rd Int. Symp. Water-Rock Interaction*, 72–74 (Alberta Research Council, Edmonton, 1980).
- Wogelius, R. A. & Walther, J. V. *Geochim. cosmochim. Acta* **55**, 943–954 (1991).
- Hewkin, D. J. & Prince, R. H. *Coord. Chem. Rev.* **5**, 45–73 (1970).

ACKNOWLEDGEMENTS. We thank J. Walther for comments. This research was supported by grants from the NSF, the Petroleum Research Fund as administered by the American Chemical Society, and the US Department of Energy under the Basic Energy Sciences program.

In the second phase, tension is developed in the maggot's body wall to power the jump. With the mouth-hooks firmly anchored posteriorly, commencement of helical-muscle contraction causes the maggot's body to buckle inwards on the inside of the bend at the junction between abdominal segments 5 and 6 (Fig. 2c). At peak muscle contraction the loop becomes flattened towards the ground (Fig. 2c). The maggot's body is now held in a tense, bulging and mechanically unstable state owing to a lack of longitudinal tension on the compressed and buckled inside-bend, and an increased tension on the stretched outside-bend (like a bent sausage-shaped balloon). As a result, elastic strain energy is stored in the maggot's cuticle, particularly around abdominal segments 4 and 5 (Fig. 2c).

Finally, a wave of muscle contraction travels anteriorly from segment 8. This causes further bulging of body segments and, on reaching the head, disengages the mouth hooks, thereby releasing the strain energy in the form of an elastic recoil. A rapid accelerated straightening of the maggot's body (the force acting over a distance of 0.8 mm) thrusts the tail against the substrate, thereby generating the lift that throws the maggot forwards into the air (Fig. 3).

Thus, although most pressurized cylindrical animals avoid kinking and bulging², the fruit-fly larva turns a potentially dangerous mechanical instability² into an elegant jumping mechanism. Whereas crawling limits maggots to speeds that are well below their own body length (BL) per second (0.2 BL s^{-1} ;

TABLE 1 Mechanical performance of four animals in standing jumps

	Flea (<i>Pulex</i>)*	Fruit-fly larva (<i>Ceratitis</i>)	Click beetle (<i>Athous</i>)†	Locust (<i>Schistocerca</i>)‡
Body mass (mg)	0.49	17	40	2000
Height of jump (cm)	20	7	30	30
Jump distance (cm)	39	12	0	80
Acceleration distance (mm)	0.8	0.8	0.65	40
Acceleration or take-off time (ms)	0.79	1.37	0.64	25
Acceleration relative to gravity (g)	245	86	320	14.5
Take-off speed (m s^{-1})	1.98	1.17	2.4	3.4
Take-off angle	50°	60°	90°	55°
Force for take-off (g)	0.12	1.5	15	29

Table compiled after * ref. 10; † ref. 6; and ‡ ref. 11.

Fig. 3), a single jump increases their speed by two orders of magnitude to 59 BL s^{-1} (Fig. 3).

Fruit-fly larvae are particularly vulnerable to parasitization and predation when they leave their host fruit to pupate in the ground^{3,4}. In my garden, on 6 out of 9 encounters, ants (*Tetramorium* sp.) attacking *Ceratitis* larvae were thrown off

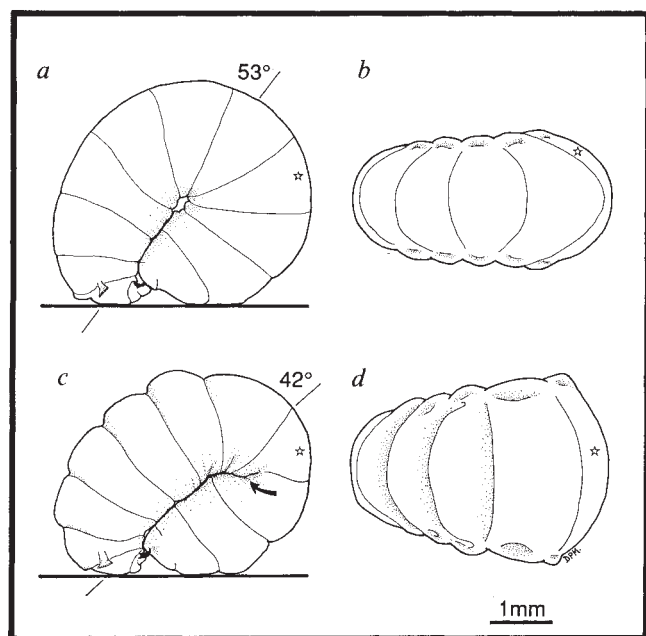


FIG. 2 Lateral (a, c) and dorsal (b, d) profiles (traced from 35 mm photographic slides) of the two phases of jump preparation in *Ceratitis*. a and b, End of stage one: the maggot is standing in an unusually tight unbuckled loop whose radius of curvature is only 1.12 times the maggot's own body radius¹⁵. The loop's long axis is inclined backwards at an angle of about 53° to the horizontal (diagonal line). c and d, End of stage two: note the dorsal bulging of thoracic and abdominal segments 1 to 3 in contrast to the stretched dorsal surface of segments 4 and 5. The abdomen is buckled inwards along the junction between segments 5 and 6 (c, curved arrow). Segment 5 (c, star) becomes 'pinched off' and compressed between segments 4 and 6. As a result, segments 4 and 5 are forced to bulge out laterally (d). The looped body is inclined at about 42° to the horizontal (c, diagonal line). Since loop formation compresses the muscles located on the inside of the loop, helical muscle contraction exerts its greatest effect circumferentially. From video analysis, stage 1 of jump preparation (loop formation) takes 0.9 s (by deduction), and stage 2 (tension generation) takes 0.7 s (± 0.1 ; $n=10$). The total jump-preparation time is 1.6 s (± 0.2 ; $n=10$).

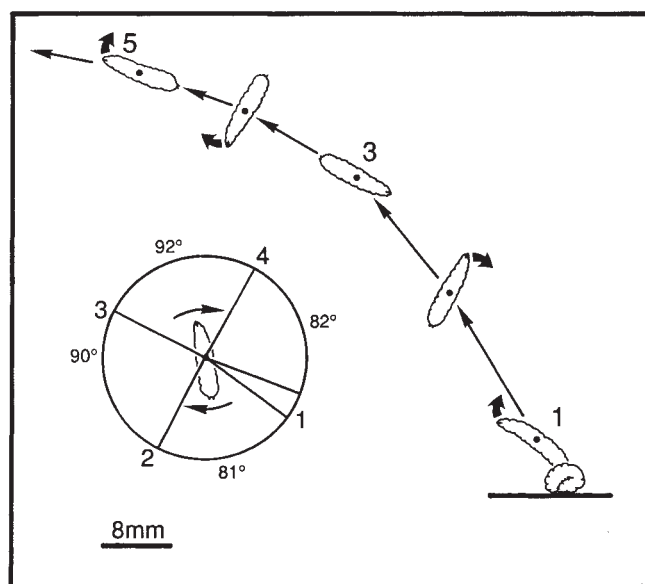


FIG. 3 The trajectory of a jumping *Ceratitis capitata* larva filmed on video with a VHS camcorder fitted with an internal shutter operating at 1 ms. The interval between successive frames is 40 ms. On leaving the ground at an angle of 60° (± 3.7 ; $n=30$) to the horizontal, the maggot spins at 42 rad s^{-1} (± 8.5 ; $n=10$) in a backward somersault. The inset shows the angle of rotation for each of the 4 intervals shown. The backward spin results from the head being flung upwards and backwards while the tail is simultaneously thrust against the substrate. The forwards velocity component results from the asymmetric loop formation and body compression which places the tip of the abdomen anterior to the point of rotation around which the maggot's body uncurls (Fig. 2c, curved arrow). During a typical jump, the maggot reaches a height of 7 cm (± 0.5 ; $n=18$; maximum=8.2 cm) and covers a horizontal distance of 12 cm (± 0.8 ; $n=63$; maximum=16 cm) in 0.24 s before touching the ground. This represents an average velocity of 0.5 m s^{-1} per jump. On landing, the maggot bounces randomly and crawls between jumps, thus slowing down the overall rate of locomotion over greater distances. At 24 °C, maggots jumped 8 times per minute, covering a total distance of 116 cm (± 17 ; $n=20$), at an average rate of locomotion of 1.9 cm s^{-1} . Crawling maggots manage, at best, 0.2 cm s^{-1} (± 0.06 ; $n=12$) on a damp smooth wood surface.

during the maggot's explosive jump. The ants subsequently failed to relocate their prey which landed several centimetres away. Thus, the unpredictable nature and force of jumping may prove to be an effective escape mechanism for fruit-fly larvae.

Given their total lack of appendages and their soft hydraulic-based skeletal system, it is surprising that maggots jump⁵. Indeed, jumping maggots appear to be the only known examples of jumping by soft-bodied legless organisms. Other examples of legless jumping are found in hard-bodied legged invertebrates, but these employ alternative mechanisms to jump (click beetles⁶, springtails⁷, bristletails⁸ and millipedes⁹). The jumping performances of maggots and click beetles, together with two other well known jumpers, are compared in Table 1.

In light of the ability of these maggots to jump, a caterpillar-like body plan which uses hydraulic-based skeletal and locomotory systems obviously does not preclude high-speed locomotion¹. In addition, the reported restrictions on stride length¹ are overcome to a degree by caterpillars (Geometridae) and leeches looping instead of crawling. Clearly, the hydraulic body of caterpillar-like organisms is more versatile than was previously believed. □

Received 27 August; accepted 9 October 1991.

- Casey, T. M. *Science* **252**, 112-114 (1991).
- Wainwright, S. A. *Axis and Circumference: The Cylindrical Shape of Plants and Animals* (Harvard University Press, Cambridge, Massachusetts, 1988).
- Wong, T. T. Y., Ramadan, M. M., McInnis, D. O. & Mochizuki, N. *J. econ. Entomol.* **83**, 779-783 (1990).
- Wong, T. T. Y., Mochizuki, N. & Nishimoto, J. I. *Environ. Entomol.* **13**, 140-145 (1984).
- Wille, J. *Zool. Jb. Allg. Zool.* **39**, 301-320 (1922).
- Evans, M. E. G. *J. Zool.* **167**, 319-336 (1972).
- Manton, S. M. *The Arthropods* (Clarendon, Oxford, 1977).
- Evans, M. E. G. *J. Zool.* **176**, 49-65 (1975).
- Evans, M. E. G. & Blower, J. G. *Nature* **246**, 427-428 (1973).
- Schmidt-Nielsen, K. *Animal Physiology: Adaptation and Environment* 4th edn (Cambridge University Press, UK, 1990).
- Alexander, R. McN. *Animal Mechanics* 2nd edn (Blackwell, Oxford, 1983).
- Phillips, V. T. *Memoirs of the American Entomological Society* No. 12 (Philadelphia, 1946).
- Shadwick, R. E. & Gosline, J. M. *J. exp. Biol.* **114**, 259-284 (1985).
- Vogel, S. *Life's Devices: The Physical World of Animals and Plants* (Princeton University Press, NJ, 1988).
- Alexander, R. McN. *J. theor. Biol.* **124**, 97-110 (1987).

ACKNOWLEDGEMENTS. I thank P. Maitland for assistance, H. Benze for translations, D. Dawson for the loan of video equipment and M. Villet for identifying ants.

Self-organizing neural network that discovers surfaces in random-dot stereograms

Suzanna Becker & Geoffrey E. Hinton

Department of Computer Science, University of Toronto,
10 King's College Road, Toronto M5S 1A4, Canada

THE standard form of back-propagation learning¹ is implausible as a model of perceptual learning because it requires an external teacher to specify the desired output of the network. We show how the external teacher can be replaced by internally derived teaching signals. These signals are generated by using the assumption that different parts of the perceptual input have common causes in the external world. Small modules that look at separate but related parts of the perceptual input discover these common causes by striving to produce outputs that agree with each other (Fig. 1a). The modules may look at different modalities (such as vision and touch), or the same modality at different times (for example, the consecutive two-dimensional views of a rotating three-dimensional object), or even spatially adjacent parts of the same image. Our simulations show that when our learning procedure is applied to adjacent patches of two-dimensional images, it allows a neural network that has no prior knowledge of the third dimension to discover depth in random dot stereograms of curved surfaces.

The simplest way to get the outputs of two modules to agree is to use the squared difference between the outputs as a cost function, and to adjust the weights (connection strengths) in each module so as to minimize this cost. Unfortunately, this usually causes each module to produce the same constant output that is unaffected by the input to the module and therefore conveys no information about it. What we want is for the outputs

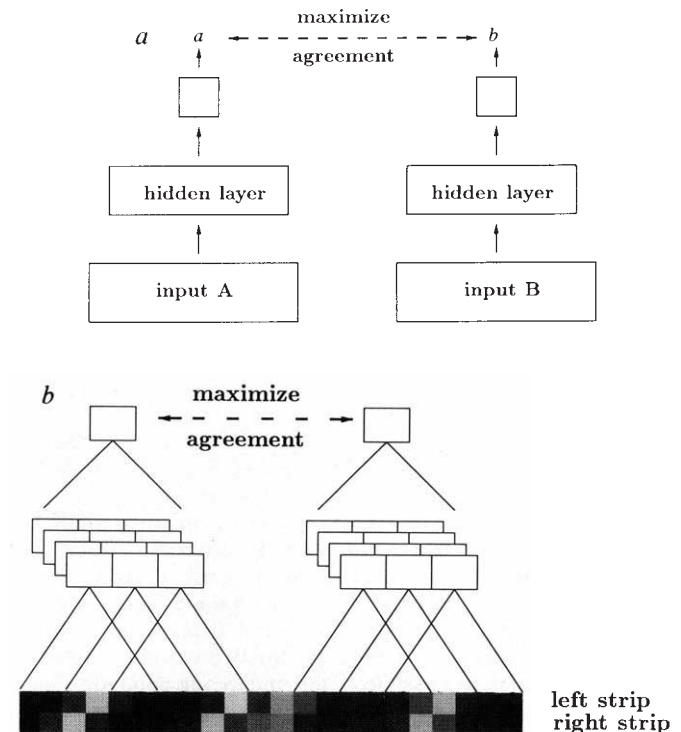
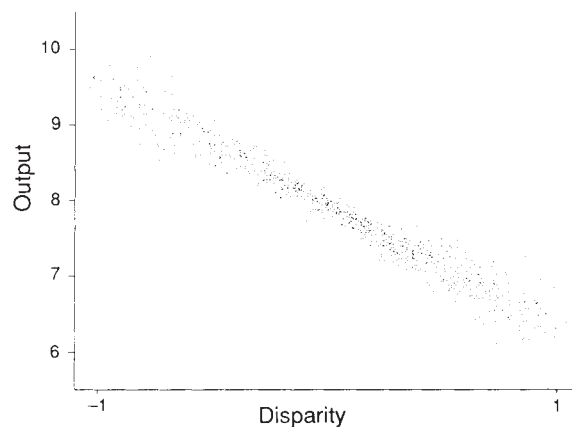


FIG. 1 a, Two modules that look at separate but related parts of the perceptual input and attempt to produce the same output. By differentiating some measure of the agreement between the outputs, we can generate a teaching signal that can be back-propagated through each module to compute how to change the connection strengths. In our initial simulations the output units are linear and the hidden units use the logistic nonlinearity $y = (1 + \exp(-\mathbf{w} \cdot \mathbf{s} - b))^{-1}$ where $\mathbf{w}$ is the weight vector of a unit, $\mathbf{s}$ is the input vector and b is the bias. b, Two modules that receive input from corresponding parts of stereo images. The input pattern is an example of a random-dot stereogram of a surface which is curved in depth. A strip of curved surface is generated by fitting a smooth curve (a cubic spline) to four or more control points whose depths are chosen at random. Random dots are scattered sparsely on the surface strip, and a pair of stereo images is made by taking two slightly different projections. The projections are filtered through a gaussian and sampled at evenly spaced sample points. In our images, disparity ranges continuously from -1 to $+1$ image pixels. The sample values in corresponding patches of the two images are used as the inputs to a module. For simplicity, we use image strips one pixel high. Each module has one layer of hidden units, organized into clusters of three feature detectors (units in the same cluster are shown in the same plane). Within a cluster, the feature detectors have the same weights but the receptive fields of neighbouring units (indicated by triangles) are translated by two pixels. Because the three hidden units within a plane are constrained to learn translated versions of the same feature, we have four planes to allow four different features to be learned. The learning algorithm adjusts the weights in each module to maximize the agreement measure, l . The derivative of l provides error signals that are propagated backwards² through the layers of each module. After each weight update, we average corresponding weights in all the modules to enforce the constraint that every module computes exactly the same function of its input vector. We also average corresponding weights of units in a cluster of three feature detectors. These equality constraints reduce the number of free parameters that must be learned which speeds the learning and limits the ability of the system to discover spurious correlations in the data that are caused by the limited size of the training set. To speed the learning further, we used a conjugate gradient optimization method for updating the weights in the modules.

FIG. 2 The output of a module as a function of the depth for 1,000 test cases of random-dot stereograms of curves surfaces. Pairs of corresponding hidden units were trained to maximize agreement using 20 iterations of steepest descent learning. Then 10 conjugate gradient iterations (typically about 250 function evaluations) were used to maximize agreement between the outputs of neighbouring modules.



of two modules to agree closely (that is, to have a small expected squared difference) relative to how much they both vary as the input is varied. When this happens, the two modules must be responding to something that is common to their two inputs. In the special case when the outputs, a and b , of the two modules are scalars, a good measure of agreement is

$$I = 0.5 \log \frac{V(a+b)}{V(a-b)} \quad (1)$$

where V is the variance over the training cases. If a and b are both versions of the same underlying gaussian signal that have been corrupted by independent gaussian noise, it can be shown that I is the mutual information between the underlying signal and the average of a and b (S.B. and G.E.H., unpublished results). By maximizing I we force the two modules to extract as pure a version as possible of the underlying common signal.

Just as back-propagation can be viewed as a multilayer nonlinear extension of linear regression because it minimizes a squared error measure, the procedure we present here can be seen as a multilayer nonlinear extension of statistical methods such as canonical correlation or alternating conditional expectation² because it maximizes the normalized covariation between the outputs.

In the following simulations, we used 1,000 training cases of random-dot stereo images such as those shown in Fig. 1b. The real-valued depth (relative to the plane of fixation) of each patch

of the surface gives rise to a disparity between intensity peaks in the left and right images. The disparity is the only property that is coherent across each stereo image. We trained a network that contained 10 modules using the architecture shown in Fig. 1b. Each module had one real-valued output and learned to extract depth by trying to maximize agreement with the outputs of immediately adjacent modules.

With random initial weights, the derivatives of I are tiny so learning is very slow. We therefore introduced an initial learning phase in which we trained corresponding pairs of units in the hidden layers of neighbouring modules to agree. It is impossible for these units to achieve very high agreement because an intermediate layer is required to extract depth properly. Also there is no pressure for different hidden units in the same module to discover different features. But even modest depth tuning of the hidden units makes the subsequent learning much easier for the output units. During the second phase of the learning we trained the output units of neighbouring modules to agree. Derivatives of this agreement were propagated backwards through the hidden layers to fine-tune the hidden units to be as useful as possible to the output units. Output units became accurately tuned to depth (Fig. 2).

So far, we have used a very simple model of coherence in which an underlying parameter at one location is assumed to be roughly equal to the parameter at a neighbouring location. This model is fine for the depth of fronto-parallel surfaces but it is far from the best model of slanted or curved surfaces.

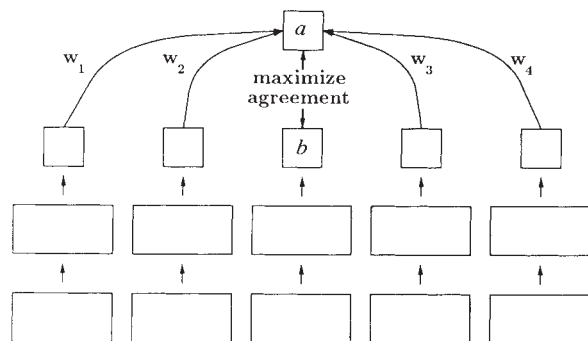


FIG. 3 The architecture of the network used for discovering locally detectable parameters that are linear combinations of nearby parameters. The network consists of multiple copies of modules like those in Fig. 2 plus a layer of interpolating units that are used for predicting the locally extracted parameter from several nearby parameters. We actually used 10 modules and the central six modules tried to maximize agreement between their outputs and contextually predicted values. We used weight averaging to constrain the interpolating function to be identical for all modules. We tested this model on several image ensembles with varying amounts of curvature, by varying the number of control points used to generate the cubic spline surfaces. After having been trained for 50 conjugate gradient iterations, the four weights learned for the interpolating function were: Two control points (CP): 0.256, 0.266, 0.258, 0.265; three CP: 0.018, 0.516, 0.531, 0.011; four CP: -0.147, 0.675, 0.656, -0.131; five CP: -0.244, 0.734, 0.746, -0.256. As the curvature increases, a characteristic pattern emerges in which positive weights are given to inputs coming from the immediately adjacent modules, and smaller negative weights are given to inputs coming from the more distant neighbours. The output of the interpolating units is similar to the response profile shown in Fig. 3, but even more finely depth-tuned. When we increase the difficulty of the learning problem by making the random-dot densities greater, the network learns a qualitatively similar interpolating function to the ones shown above but with smaller weights. This is a sensible solution because when predicting a depth from noisy estimates of nearby depths, the noise amplification is proportional to the sum of the squares of the weights.

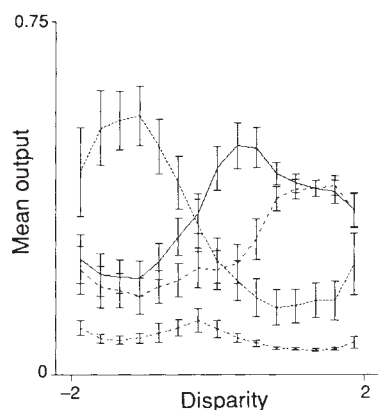


FIG. 4 The depth-tuning curves of the four competing output units of a module. The curves are averages over 1,000 test cases of random-dot stereograms of curved surfaces. The training and test patterns were created in the same manner as described in Fig. 2, but this time the disparity ranged continuously from -2 to $+2$ pixels. The agreement measure used for the learning is a modified version of the function defined in equation (1). (S.B. and G.E.H., manuscript in preparation).

Fortunately, we can use a far more general model of coherence in which the parameter at one location is assumed to be an unknown linear function of the parameters at nearby locations. The particular linear function that is appropriate can be learned by the network.

We used a network of the type shown in Fig. 3. The depth computed locally by a module, b , was compared with the depth predicted by a linear combination of the outputs of nearby modules, a , and the network tried to maximize the agreement between a and b . The contextual prediction, a , was produced by computing a weighted sum of the outputs of two adjacent modules on either side. The interpolating weights used in this sum, and all the other weights in the network, were adjusted so as to maximize agreement between locally computed and contextually predicted depths. To speed the learning, we first trained the lower layers of the network as before, so that agreement was maximized between neighbouring locally computed outputs. This made it easier to learn good interpolating weights. When the network was trained on stereograms of cubic surfaces, it learned interpolating weights of -0.147 , 0.675 , 0.656 , -0.131 . Given noise-free estimates of local depth, the optimal linear interpolator for a cubic surface is -0.167 , 0.667 , 0.667 , -0.167 .

There are many possible variations of this learning procedure. We assumed that the depth of a patch is represented by the activity level of a single linear unit. An alternative representation, which seems to be used by binocular cortical cells³, is to have a population of units each of which is tuned to a range of depths. This representation has a number of advantages, including the ability to represent uncertainty by uniformity of activity across the population. Our measure of agreement can be modified to produce such population codes (Fig. 4).

We have described the learning procedure for modules which each have a single real-valued output. For modules with several real-valued outputs, the natural way to generalize the objective function is to replace the variance by the determinant of the covariance matrix. When this version of the procedure is applied to an ensemble of images of the same two-dimensional shape with different poses (that is, different positions, sizes and orientations), it extracts the four parameters of the pose, because different local fragments of the object all agree on the pose of the whole object⁴.

The procedure we have described was designed to eliminate the need for an external teacher, but it may also overcome another weakness of supervised learning procedures. A global

external teaching signal causes interdependencies between all the weights in the network which leads to slow learning in large networks. If pairs of small modules can generate their own teaching signals by trying to maximize agreement, many pairs can learn in parallel without interference. Also, the outputs of modules that have already learned can become the inputs to other modules that look for more complex or longer-range coherence. This should make learning much faster in very large networks. □

Received 10 May; accepted 4 October 1991.

1. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Nature* **323**, 533–536 (1986).
2. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* (Chapman and Hall, London, 1990).
3. Lehigh, S. R. & Sejnowski, T. J. *J. Neurosci.* **10**, 2281–2299 (1990).
4. Zemel, R. S. & Hinton, G. E. in *Advances in Neural Information Processing Systems* Vol. 3 (eds Lippman, R. P., Moody, J. E. & Touretzky, D. S.) 299–305 (Morgan Kaufmann, San Mateo, CA, 1991).

ACKNOWLEDGEMENTS. We thank P. Brown, F. Crick, A. Jepson, B. Pearlmutter and M. Revow. This research was supported by the Canadian Institute for Advanced Research and the Ontario Information Technology Research Center.

Rapid-time-course miniature and evoked excitatory currents at cerebellar synapses *in situ*

R. Angus Silver, Stephen F. Traynelis & Stuart G. Cull-Candy

Department of Pharmacology, University College London, Gower Street, London WC1E 6BT, UK

NEUROTRANSMISSION from mossy fibre terminals onto cerebellar granule cells is almost certainly mediated by L-glutamate^{1,2}. By taking advantage of the small soma size, limited number of processes and short dendrite length of granule cells, we have obtained high-resolution recordings of spontaneous miniature excitatory postsynaptic currents (m.e.p.s.cs) and evoked currents in thin cerebellar slices³. Miniature currents have a similar time-course and pharmacology to evoked currents and consist of an exceptionally fast non-NMDA (*N*-methyl-D-aspartate) component (measured rise-time, $200\ \mu\text{s}$; estimated pre-filtered rise-time $<100\ \mu\text{s}$; decay time constant, $\tau = 1.0\ \text{ms}$), followed by $50\ \text{pS}$ NMDA channel openings that are directly resolvable. We could find no evidence for the recent proposal that miniature currents in granule cells are mediated solely by NMDA channels with a novel time course⁴. The non-NMDA receptor component of m.e.p.s.cs has a skewed amplitude distribution, which suggests potential complications for quantal analysis. The difference in time course between the m.e.p.s.cs reported here and other synaptic currents in the brain^{5–8} could reflect differences in synaptic function or electrotonic filtering; the relative contribution of these possibilities has yet to be established.

Mossy fibre-granule cell synaptic transmission invariably produces dual component excitatory postsynaptic currents (e.p.s.cs) (Fig. 1*a, c*). The fast component (mean $\pm$ s.e.m., $1,700 \pm 300\ \text{pS}$; $n = 22$) is reversibly blocked by the non-NMDA receptor antagonists CNQX (6-cyano-7-nitro-quinoxaline-2,3-dione) and NBQX (6-nitro-7-sulphamoyl-benzo(f)quinoxaline-2,3-dione)⁹ ($3\text{--}10\ \mu\text{M}$; $n = 9$; Fig. 1*b, d*), whereas the slow component can be reversibly and selectively inhibited by the NMDA receptor antagonists APV⁹ ($10\text{--}20\ \mu\text{M}$; $n = 14$, Fig. 1*a, b*) and 7-chlorokynurenate ($5\text{--}10\ \mu\text{M}$; $n = 6$), and by Mg^{2+} ($1\ \text{mM}$, at negative potentials). The pharmacologically isolated fast non-NMDA component (Fig. 1*a, c*, inset) had a $10\text{--}90\%$ rise-time of $410 \pm 30\ \mu\text{s}$ and decay time constant of $1.3 \pm 0.1\ \text{ms}$ (monotonically rising events, $n = 15$ cells); this contrasted with the characteristically slow rise ($10\text{--}90\%$ rise-time $9 \pm 1\ \text{ms}$; $n = 7$) and decay

($\tau = 52 \pm 5$ ms; $n = 21$) of the synaptic component carried by NMDA receptors (Fig. 1*b, d*)^{5-7,10}. To test whether intracellular dialysis associated with whole-cell recording¹¹ affected the time course of the synaptic currents, we also used the perforated-patch method¹² (Fig. 1*c, d*), in which the integrity of the perforated-patches was verified. Figure 1 shows a perforated-patch recording with an unusually low series resistance selected to facilitate comparison with whole-cell recordings. The two techniques gave similar results, and could not account for the difference between our results and previous observations⁴.

Spontaneous m.e.p.s.cs occurred at low frequency (0.21 ± 0.05 Hz; $n = 7$ cells) in some unstimulated cells bathed in $0.5 \mu\text{M}$ tetrodotoxin (TTX). As Fig. 2*a* illustrates, they typically consisted of a fast component (180 ± 30 pS; $n = 7$) followed by clearly resolvable single-channel currents. Averages of m.e.p.s.cs (Fig. 2*b*) revealed that the channel openings following the fast component constituted a second slower synaptic component ($\tau = 37 \pm 10$ ms; $n = 6$) that was sensitive to APV and Mg^{2+} ($n = 7$). Figure 2*c* illustrates the fast time course of m.e.p.s.cs recorded in APV; the 10–90% rise-time (200 ± 10 μs) and decay time constant (1.0 ± 0.1 ms; $n = 7$) were significantly faster than for evoked e.p.s.cs. In common with evoked e.p.s.cs, the fast component of m.e.p.s.cs was sensitive to CNQX (Fig. 2*c*; $n = 5$). The amplitude ratio of the NMDA to non-NMDA conductance change for m.e.p.s.cs (0.08 ± 0.01 ; $n = 6$) was not significantly different from that of the evoked current (0.12 ± 0.01 ; $n = 21$, $P > 0.05$, unpaired t -test), which is consistent both with the idea that evoked currents result from the summation of m.e.p.s.cs, and that any undetected pure NMDA miniature currents do not make a significant contribution to the evoked current. The pharmacology, time course and size of the m.e.p.s.cs all support the assumption that these spontaneous currents represent the elementary postsynaptic response to the release of a transmitter packet. We could not fit amplitude histograms of the peak non-NMDA m.e.p.s.cs with a single gaussian function, because they appeared clearly skewed (Fig. 2*d*), resembling miniature

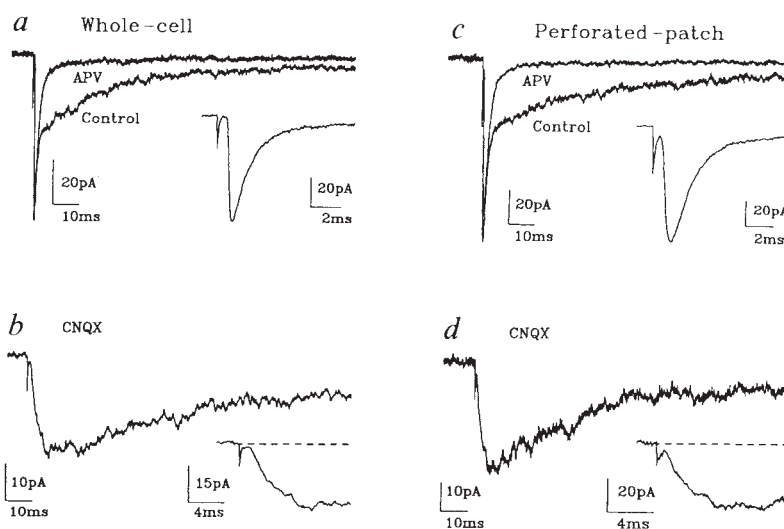
distributions at both cholinergic¹³ and GABA_A synapses^{14,15}. If the events lying outside the main distribution, whose amplitude could result from either pre- or postsynaptic factors, contribute significantly to the evoked current, then extracting the release statistics for this central glutamate synapse is likely to be complicated. This suggests that simplifying assumptions, justifiable at the neuromuscular junction¹⁶, may not be appropriate for central glutamate synapses, as previously discussed¹⁷.

Despite the small size of granule cells, the waveform of the synaptic current will still be somewhat distorted by low-pass filtering of both the electrode-cell circuit and recording apparatus. As Fig. 3*a* shows, the averaged m.e.p.s.c. rise-time was close to the rise-time of the measurement system. The averaged m.e.p.s.c. time course (interpolated from 15–50 to 90–100 kHz) could be fitted with an exponentially rising and decaying function that was passed sequentially through a digital model cell constructed with the measured cell parameters, and a gaussian filter that replicated the apparatus filtering. The m.e.p.s.c. was best fitted (Fig. 3*b*) by a function with a 10–90% rise-time of 95 ± 16 μs ($n = 7$). However, this value may represent an upper limit because the cutoff frequency of the electrode-cell circuit (typically 1.5–2 kHz) results in rise-times faster than 50 μs being filtered to similar final values (Fig. 3*c*). The estimated peak conductance underlying the m.e.p.s.c. was $113 \pm 4\%$ ($n = 7$) of the recorded value; the decay time course was unchanged. Figure 3*a* also compares the time course of the m.e.p.s.c. with the apparently monotonically rising evoked non-NMDA e.p.s.c. The slower rise of the e.p.s.c. may well result from asynchronous release of transmitter.

Analysis of NMDA channel current amplitudes in the tail of spontaneous currents (Fig. 2*a*) revealed that channel openings of 48 ± 1 pS ($n = 4$ cells) underlie the slow synaptic component, although some lower conductance levels were also present. This similarity between the main conductance state of synaptic and somatic¹⁸ NMDA channels has also been noted in cultured hippocampal neurons¹⁹. Our mean value for the peak NMDA

FIG. 1 Pharmacology and time course of evoked e.p.s.cs. *a* and *b*, Whole-cell recording at -60 mV; *c* and *d*, perforated-patch recording ($R_s = 60$ M Ω) at -65 mV. *a*, Control: e.p.s.cs (average of 22) could be fitted (not shown) by the sum of 2 exponential functions ($\tau_{\text{non-NMDA}} = 1.5 \pm 0.1$ ms; $\tau_{\text{NMDA}} = 52 \pm 5$ ms; $A_{\text{NMDA}}/A_{\text{non-NMDA}} = 0.12 \pm 0.01$, where A is the current component amplitude; $n = 22$ cells). APV: average of 19 events in $20 \mu\text{M}$ APV (2-amino-5-phosphonopentanoic acid). Inset shows the non-NMDA synaptic component on a faster timescale. *b*, Average of 23 e.p.s.cs in $10 \mu\text{M}$ CNQX possess only a slow NMDA component. *c*, Control: e.p.s.cs (average of 20 events) measured using the perforated-patch method ($\tau_{\text{non-NMDA}} = 2.0 \pm 0.5$ ms; $\tau_{\text{NMDA}} = 49 \pm 5$ ms; $A_{\text{NMDA}}/A_{\text{non-NMDA}} = 0.15 \pm 0.02$; $n = 4$ cells). APV: e.p.s.cs (average of 44) in $20 \mu\text{M}$ APV (10–90% rise-time = 500 ± 100 μs ; decay time constant $\tau = 2.1 \pm 0.4$ ms; $n = 4$ cells). *d*, Perforated-patch recording of NMDA synaptic component (average of 21) in $10 \mu\text{M}$ CNQX (10–90% rise-time = 7.2 ± 0.7 ms; $\tau = 51 \pm 1$ ms; $n = 3$). The current of the non-NMDA component was linearly related to voltage, reversing at -2 ± 2 mV ($n = 4$); the decay time constant seemed to be independent of voltage between -80 mV and $+40$ mV. Nystatin (200 – $300 \mu\text{g ml}^{-1}$) and fluorescein or rhodamine isothiocyanate-conjugated dextran (1.5 mM) were added to the pipette solution for the perforated-patch technique. Exclusion of the high-molecular-mass fluorescent dye from the cell interior verified the integrity of the perforated patch at the end of each recording; the results from all but about 7% of experiments were discarded because the nystatin patch ruptured spontaneously during the experiment.

METHODS. Thin cerebellar slices (150 – $180 \mu\text{m}$) from 11–17-day-old Sprague-Dawley rats were viewed under Nomarski optics (22 – 25°C)³. External solutions (Mg^{2+} -free Krebs solution bubbled with 95% O_2 and 5% CO_2) contained glycine (3 – $5 \mu\text{M}$ in control, 3 – $10 \mu\text{M}$ in CNQX) and 10 – $20 \mu\text{M}$ bicuculline methobromide. Patch pipettes of thick-wall borosilicate glass, coated with Sylgard resin, were polished to 5 – 15 M Ω , and filled with (mM):



110 CsF; 30 CsCl; 4 NaCl; 0.5 CaCl₂; 10 HEPES; 5 EGTA (pH 7.3). Cell capacitance (3.4 ± 0.2 pF, $n = 29$ cells). R_s (whole-cell configuration, 29 ± 2 M Ω ; $n = 29$; nystatin patch, 130 ± 30 M Ω ; $n = 4$), and input resistance (4.7 ± 0.8 G Ω ; $n = 21$), were measured from both the List amplifier and the transient capacitance currents produced by 2 – 5 mV voltage steps digitized at 100 – 125 kHz after 20 – 40 kHz filtering (all reported frequencies ~ 3 dB). The mossy fibre pathway was stimulated through a second patch electrode placed in the surrounding tissue; transmission could be evoked in $\sim 20\%$ of granule cells. Evoked e.p.s.cs and m.e.p.s.cs were low-pass filtered (Barr and Stroud, Bessel-type) to 2 – 10 kHz and digitized on line or off FM tape (bandwidth d.c. to 5 – 7.5 kHz) at 10 – 100 kHz.

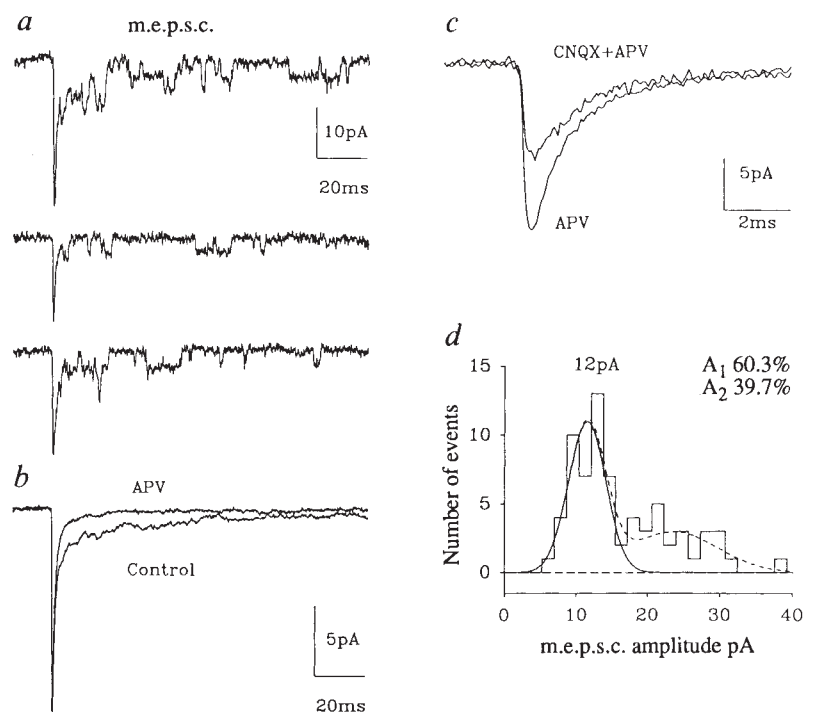
component of the m.e.p.s.c. was 20 ± 6 pS ($n = 6$ cells). Assuming that the NMDA synaptic current is mainly carried by receptor activations that generate prolonged clusters (>1 burst) of openings (if 26% of activations are prolonged clusters and open probability within a cluster is 0.6)²⁰ and that only 50% of these receptors are available at physiological pH²¹, then this current indicates there could be as few as five NMDA receptors in the area of postsynaptic membrane exposed to the single transmitter packet.

There appear to be two possible explanations for the difference between the properties of the m.e.p.s.c.s that we describe and previous observations in which it was concluded that the m.e.p.s.c. was carried solely by NMDA receptors with a novel time course⁴. First, our modelling suggests that attenuation of the peak amplitude by electrode series resistance (R_s) filtering is severe for fast decaying events ($\tau \leq 1$ ms). The filtering imposed by the high R_s (70–130 M Ω) of the perforated-patch method would attenuate m.e.p.s.c.s by 27–41% (compared with 12% here) and thereby hamper identification of the non-NMDA component of m.e.p.s.c.s, given their small amplitude. Second, in experiments performed in the presence of non-NMDA receptor blockers, we could not discriminate convincingly between spontaneous synaptic NMDA receptor activation and the 'tonic' background NMDA channel activity present in cells in thin slices³¹. That is, m.e.p.s.c.s could be identified reliably only when the fast non-NMDA component was present. Indeed, if bursts of 50-pS openings (without a distinguishing non-NMDA component) were misclassified as synaptic in origin, then averaging such currents by aligning their onset will yield a spuriously fast rise-time and a decay time constant equal to the mean burst-length of the NMDA channels (6–11 ms in granule cells^{18,22}). These values are similar to those previously reported for m.e.p.s.c.s at this synapse⁴.

Several clear implications about transmission at this and other central glutamate synapses can be drawn from our experiments. Spontaneous m.e.p.s.c.s which have both non-NMDA and NMDA receptor components represent the elementary postsynaptic response to the release of a transmitter packet. Resolution of channel openings in m.e.p.s.c.s suggests that comparatively few active NMDA receptors are present in the postsynaptic membrane¹⁹. The similar time course and pharmacology of m.e.p.s.c.s and evoked e.p.s.c.s indicates that spontaneous transmission at the mossy fibre-granule cell synapse is conventional in that it is mediated by both classes of glutamate receptor and not solely by NMDA receptors⁴. Important questions such as how m.e.p.s.c. size is determined and whether simplifying assumptions are appropriate for quantal analysis of glutamate synapses^{15,17} are raised by the skewed m.e.p.s.c. amplitude distribution. The data also suggest that activation kinetics cannot be easily extracted even from apparently monotonically rising evoked e.p.s.c.s because the rising phase is slowed, presumably by asynchrony of transmitter release. However, our estimate of <100 μ s for the 10–90% non-NMDA receptor activation time in m.e.p.s.c.s (at room temperature) should be unaffected by this problem. Both the rapid rise and decay of the synaptic non-NMDA component in granule cells is appreciably faster than reported for other brain synapses^{5–8} and may be comparable to some spinal synapses^{23,24}. Moreover, the decay time of the non-NMDA m.e.p.s.c. component is consistent with individual channel open times recorded from these cells²². Even in recordings where particular attention has been paid to the problems of voltage nonuniformity and filtering, such as the climbing fibre-Purkinje cell synapse⁸, the time course of the non-NMDA synaptic current decay exceeds by 6-fold that in the granule cells. Possible explanations for the marked difference in synaptic current time course, aside from electrotonic filtering, include

FIG. 2 Miniature e.p.s.c.s in cerebellar granule neurons. *a*, Three examples of spontaneous m.e.p.s.c.s consisting of an initial fast non-NMDA component followed by individually resolved ~ 50 pS NMDA channel openings (same cell, $V_m = -60$ mV; overall filtering 1 kHz). *b*, Control: average of 115 m.e.p.s.c.s from the same cell; the trace could be fitted with two exponential functions ($\tau_1 = 1.1$ ms; $\tau_2 = 26$ ms; $A_{\text{non-NMDA}} = 12.6$ pA; $A_{\text{NMDA}} = 2.7$ pA). APV: average of 70 m.e.p.s.c.s from the same cell in 10 μ M APV, which abolished the NMDA component and had no effect on frequency ($n = 7$ cells). *c*, APV: same averaged non-NMDA current component of m.e.p.s.c.s as in *b*, now 3-kHz filtering and faster time-scale to illustrate rapid time course (10–90% rise-time = 142 μ s; $\tau = 1.1$ ms). CNQX + APV: average of 22 events in the presence of APV (10 μ M) and a submaximal concentration of CNQX (0.5 μ M). The average non-NMDA current present in CNQX will be an overestimate as some partially antagonized m.e.p.s.c.s will be lost in the baseline noise. *d*, Skewed amplitude histogram of the peak non-NMDA m.e.p.s.c.s (-60 mV) for 70 events in 10 μ M APV (bin width, 1.7 pA) is not well fitted by a single gaussian function. If the distributions are multi-modal¹³, fitting the data with two components (dashed line) allows an estimate of the main distribution (solid line) and the proportion of events lying outside (means, 117 ± 24 pS; area 1 (A_1), 75% and 247 ± 66 pS; A_2 , 25%; $n = 4$, 66–189 events per distribution).

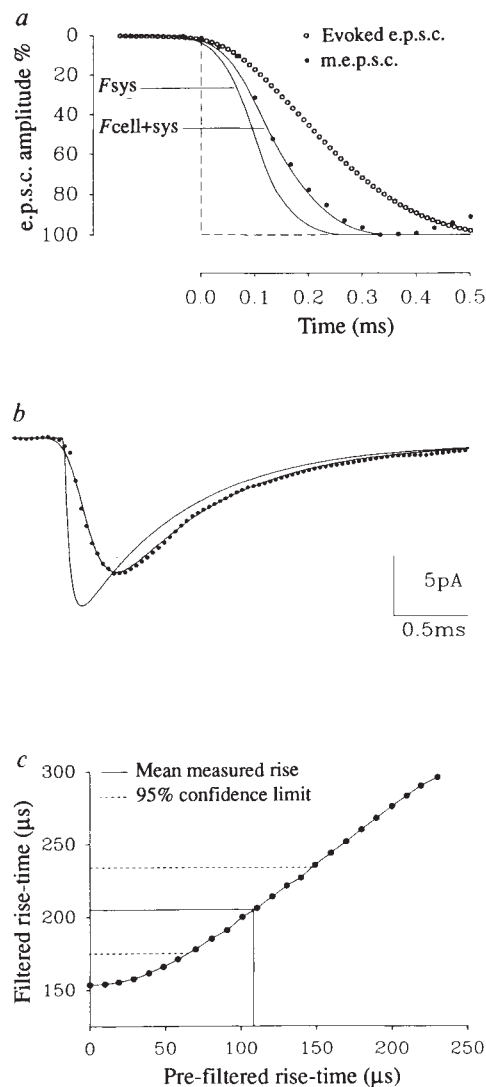
METHODS. The following criteria were used to identify m.e.p.s.c.s (10 cells; 0.5 μ M TTX): (1) inward currents were asymmetrical with a monotonic fast rise and approximately exponential decay; (2) the mean current of the event was significantly different from an equivalent length of pre-event baseline; (3) event amplitudes less than 4–6 pA (depending on baseline variance) were ignored. Average time course for evoked e.p.s.c.s and spontaneous m.e.p.s.c.s was produced by aligning the first digitized point associated with the rise (usually the maximum of the derivative). The 10–90% rise-time was calculated from interpolation, or by linear regression over the appropriate data points. Peak non-NMDA current amplitudes were obtained by extrapolating the fitted exponential function to the peak time, or by



averaging the 3–10 data points at the peak. Evoked e.p.s.c. and m.e.p.s.c. decays were fitted to the sum of 1 or 2 exponential components (simplex algorithm; least-squares criterion). The ratio of non-NMDA to NMDA amplitudes was calculated from control events using the filter-corrected non-NMDA component (assuming negligible NMDA current in the first 200–400 μ s) and the amplitude at the 100% point of the NMDA component rise. Amplitude distributions were fitted using the maximum likelihood method²⁷.

FIG. 3 Rise-times of evoked e.p.s.cs and spontaneous m.e.p.s.cs. *a*, Filled circles, normalized average time course of m.e.p.s.c. rise (interpolated from 15 to 30 kHz sampling; $n=90$ m.e.p.s.cs; 10–90% rise-time = 180 μ s); open circles, normalized average time course of evoked e.p.s.c. rise (10–90% rise-time = 350 μ s). F_{sys} : time course of an instantaneously rising and exponentially decaying function ($\tau = \text{m.e.p.s.c. time constant}$) subject to the simulated system filtering (10–90% rise-time = 130 μ s). $F_{\text{sys}} + \text{cell}$: time course of the same function now passed through a two-compartment model cell (see below) and subjected to the simulated system filtering (10–90% rise-time = 180 μ s). On average, there was no significant difference (mean = 36 ± 16 μ s; $n=7$ cells, paired t -test) between the m.e.p.s.c. rise-time and that for a filtered instantaneously rising conductance change. *b*, Filled circles, average m.e.p.s.c. time course (same cell as *a*); the solid line overlying the data is the least-squares fit of an exponentially rising, exponentially decaying conductance change ($A[\exp(-t/\tau_{\text{decay}}) - \exp(-t/\tau_{\text{rise}})] + \text{offset}$) subject to the modelled cell and system filtering. The faster rising trace is the pre-filtered current predicted at the postsynaptic site assuming the holding potential at the dendrite. *c*, Plot of pre-filtered 10–90% rise-time against the filtered rise-time (modelled cell plus apparatus filtering; average of R_s , R_m , C_m , τ_{rise} , τ_{decay} and F_{sys} from the 7 cells used for m.e.p.s.c. reconstruction) for an exponentially rising and decaying event. The flattening of the curve below 50 μ s is a result of the frequency response of the cell, and highlights the difficulty of extracting information about rapid current rise-times.

METHODS. The filtering properties of granule cells were characterized by examining their electrotonic structure both experimentally and using digital models²⁸ under voltage- and current-clamp based on anatomical (M. Farrant, R.A.S. and S.G.C.-C.)²⁹ and electrical measurements. Time constants for charging curves measured under both current-clamp (0.5–10 pA steps) and voltage-clamp (see methods for ionic conditions) were used to calculate the electrotonic length ($L = 0.05 \pm 0.01$; $n=7$) and the dendritic dominance ($\rho = 0.98 \pm 0.01$; $n=6$)³⁰. These values indicate that granule cells are electrically extremely compact and compare well with $L=0.02$ and $\rho=1.16$ predicted for our multicompartiment model (assuming 100 Ω cm intracellular resistivity) based on the measured dimensions of 3.55 μ m soma radius, 13.5 μ m dendritic length, 0.5 μ m dendritic diameter and 4 dendrites per cell ($n=151$ cells); 4 'digits' per dendrite were also included²⁹. This geometry together with the measured capacitance and input resistance give a specific membrane capacitance of $1.0 \mu\text{F cm}^{-2}$ and resistance of $16.1 \text{ k}\Omega \text{ cm}^2$. Current-clamp modelling of voltage attenuation as a function of distance and frequency (0.1–10 kHz) further indicates that these cells can be considered as one electrical compartment. Miniature e.p.s.c. was reconstructed using a model under voltage-clamp with measured parameters from each individual cell. There was no difference between a single (cell) or sequentially implemented (dendrite and then cell) digital RC filter(s) (cutoff frequency for mean cell size was 1.57 kHz, -3 dB , $R_s = 30 \text{ M}\Omega$), which is also consistent with a single electrical compartment. System filtering (amplifier, filter, tape recorder) was approximated with a digital gaussian filter with a cutoff frequency (4.91 kHz, -3 dB) chosen to filter a step-response function to the same extent as our apparatus.



differences in both the time course of transmitter release and kinetic properties of synaptic receptors²⁴. The latter proposal is compatible with the differential regional expression of non-NMDA receptor subunits^{25,26}.

Note added in proof: Evoked currents with a fast rise-time and a mean decay time constant of 2.4 ms have recently been found in non-pyramidal neurons in visual cortex³². □

Received 24 June; accepted 9 October 1991.

- Somogyi, P., Halasy, K., Somogyi, J., Storm-Mathisen, J. & Ottersen, O. P. *Neuroscience* **19**, 1045–1050 (1986).
- Ji, Z., Aas, J.-E., Laake, J., Walberg, F. & Ottersen, O. P. *J. comp. Neurol.* **307**, 296–310 (1991).
- Edwards, F. A., Konnerth, A., Sakmann, B. & Takahashi, T. *Pflügers Arch. ges. Physiol.* **414**, 600–612 (1989).
- D'Angelo, E., Rossi, P. & Garthwaite, J. *Nature* **346**, 467–470 (1990).
- Forsythe, I. D. & Westbrook, G. L. *J. Physiol.* **396**, 515–533 (1988).
- Hestrin, S., Nicoll, R. A., Perkel, D. J. & Sah, P. *J. Physiol.* **422**, 203–225 (1990).
- Keller, B. U., Konnerth, A. & Yaari, Y. *J. Physiol.* **435**, 275–293 (1991).
- Llano, I., Marty, A., Armstrong, C. M. & Konnerth, A. *J. Physiol.* **434**, 183–213 (1991).
- Watkins, J. C., Krogsgaard-Larsen, P. & Honore, T. *Trends pharmacol. Sci.* **11**, 25–33 (1990).
- Lester, R. A. J., Clements, J. D., Westbrook, G. L. & Jahr, C. E. *Nature* **346**, 565–567 (1990).
- Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. *Pflügers Arch. ges. Physiol.* **391**, 85–100 (1981).
- Horn, R. & Marty, A. *J. gen. Physiol.* **92**, 145–159 (1988).
- Bornstein, J. C. *J. Physiol.* **282**, 375–398 (1978).
- Robert, N., Miles, R. & Korn, H. *J. Physiol.* **428**, 707–722 (1990).
- Edwards, F. A., Konnerth, A. & Sakmann, B. *J. Physiol.* **430**, 213–249 (1990).
- Katz, B. in *The Release of Neural Transmitter Substances* (Liverpool University Press, UK, 1969).

- Larkman, A., Stratford, K. & Jack, J. *Nature* **350**, 344–347 (1991).
- Cull-Candy, S. G., Howe, J. R. & Ogden, D. C. *J. Physiol.* **400**, 189–222 (1988).
- Robinson, H. P. C., Sahara, Y. & Kawai, N. *Biophys. J.* **59**, 295–304 (1991).
- Gibb, A. J. & Colquhoun, D. *Proc. R. Soc. Lond. B* **243**, 39–45 (1991).
- Traynelis, S. F. & Cull-Candy, S. G. *J. Physiol.* **433**, 727–763 (1991).
- Howe, J. R., Cull-Candy, S. G. & Colquhoun, D. *J. Physiol.* **432**, 143–202 (1991).
- Finkel, A. S. & Redman, S. J. *J. Physiol.* **342**, 615–632 (1983).
- Nelson, P. G., Pun, R. Y. K. & Westbrook, G. L. *J. Physiol.* **372**, 169–190 (1986).
- Keinonen, K. et al. *Science* **249**, 556–560 (1990).
- Boulter, J. et al. *Science* **249**, 1033–1037 (1990).
- Colquhoun, D. & Sigworth, F. J. in *Single Channel Recording* (eds Sakmann, B. & Neher, E.) 191–263 (Plenum, New York, 1983).
- MacGregor, R. J. *Neural and Brain Modelling* (Academic, San Diego, 1987).
- Jakab, R. L. & Hamori, J. *Anat. Embryol.* **179**, 81–88 (1988).
- Rall, W. *Biophys. J.* **9**, 1483–1508 (1969).
- Sah, P., Hestrin, S. & Nicoll, R. A. *Science* **246**, 815–818 (1989).
- Stern, P., Edwards, F. A. & Sakmann, B. *J. Physiol.* (in the press).

ACKNOWLEDGEMENTS. This work was supported by the Wellcome Trust and the NIH. We thank M. Farrant for his help, D. Colquhoun, F. Edwards and A. Gibb for discussion and comments on the manuscript, and T. Honore for a gift of NBQX.

Inhibition by chloroquine of a novel haem polymerase enzyme activity in malaria trophozoites

A. F. G. Slater & A. Cerami

The Picower Institute for Medical Research, 350 Community Drive, Manhasset, New York 11030, USA

THE incidence of human malaria has increased during the past 20 years; 270 million people are now estimated to be infected with the parasite¹. An important contribution to this increase has been the appearance of malaria organisms resistant to quinoline-containing antimalarials such as chloroquine and quinine². These drugs accumulate in the acid food vacuoles of the intraerythrocytic-stage malaria parasite³⁻⁵, although the mechanism of their specific toxicity in this organelle is uncertain. The primary function of the food vacuole is the proteolysis of ingested red cell haemoglobin^{6,7} to provide the growing parasite with essential amino acids. Haemoglobin breakdown in the food vacuole releases haem, which if soluble can damage biological membranes⁸ and inhibit a variety of enzymes^{9,10}. Rather than degrading or excreting the haem, the parasite has evolved a novel pathway for its detoxification by incorporating it into an insoluble crystalline material called haemozoin or malaria pigment¹¹. These crystals form in the food vacuole of the parasite concomitant with haemoglobin degradation, where they remain until the infected red cell bursts. The structure of haemozoin comprises a polymer of haems linked between the central ferric ion of one haem and a carboxylate side-group oxygen of another¹². This structure does not form spontaneously from either free haem or haemoglobin under physiological conditions¹²⁻¹⁴, and the biochemistry of its formation is unclear. Here we report the identification and characterization of a haem polymerase enzyme activity from extracts of *Plasmodium falciparum* trophozoites, and show that this enzyme is inhibited by quinoline-containing drugs such as chloroquine and quinine. This provides a possible explanation for the highly stage-specific anti-malarial properties of these drugs.

When an extract of isolated *P. falciparum* trophozoites was incubated with haem in sodium acetate at pH 4.8, a fraction of the haem was converted into a product with haemozoin-like properties (Fig. 1a). This product was resistant to hydrolysis in weakly basic solvents and was insensitive to protease digestion. A Fourier-transform infrared spectrum of the isolated product confirmed its identity with haemozoin (Fig. 2). The haemozoin specific-absorbances¹² at 1665 and 1211 cm⁻¹ were enhanced after the incubation, although the spectrum did not contain features characteristic of free haematin (for example, no broad peak centred at 1226 cm⁻¹). These results suggest that a component(s) of trophozoites can promote the polymerization of haem to form haemozoin, although this finding is complicated by the initial presence of haemozoin in the extract. To overcome this problem, the experiments were repeated with ¹⁴C-haem as the substrate, such that after liquid scintillation counting only the *de novo* synthesized haemozoin was detected. The trophozoite extract again catalysed the conversion of haem into haemozoin (Fig. 1b), whereas extracts of either uninfected red blood cells or macrophages did not. Solutions of either denatured haemoglobin or albumin (which both have a high nonspecific affinity for haem) were also negative for haem polymerase activity in this assay (data not shown). The malaria trophozoite extract could also use haemoglobin for haemozoin formation (Fig. 4b). This is significant as it is not known whether the physiological substrate in the food vacuole is haemoglobin-bound haem, free haem, or haem associated with another vacuolar haem-binding protein.

If the trophozoite extract was separated by centrifugation into a soluble protein-rich fraction and a pellet containing pre-

dominantly membranes and haemozoin, the haem polymerase activity was only found in the pellet (data not shown). The same result was obtained when trophozoites were extracted in the presence of either 0.5% CHAPS or 1% Triton X-100, although the activity was destroyed after extraction in 1% SDS (data not shown). The activity declined rapidly during storage of the pellet at -20 °C, although pretreatment of the pellet fraction with either protease E or protease K (1 mg ml⁻¹ in 50 mM Tris-HCl, pH 7.4) did not affect enzyme activity (data not shown). If the haemozoin/membrane pellet was replaced in the assay with a

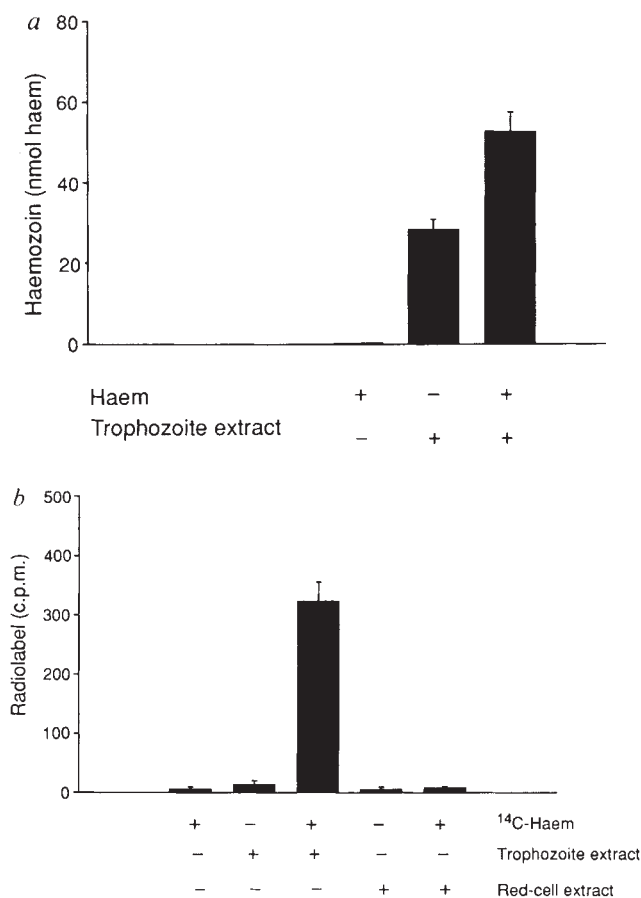


FIG. 1 Haemozoin formation in an extract of malaria trophozoites. *a*, Trophozoite extracts (0.5 mg protein) were incubated overnight at 37 °C in 500 mM sodium acetate, pH 4.8, in the presence or absence of 400 µM haematin. Results (mean ± s.e.m.) from 11 separate experiments are shown. *b*, A trophozoite extract (65 µg protein) or a red-cell extract (1.2 mg protein) was incubated for 9 h as above in the presence or absence of 140 µM ¹⁴C-haematin. Incorporation of radiolabel into haemozoin was determined in triplicate (mean ± s.e.m.).

METHODS. *P. falciparum* clone HB-3 was cultured in A⁺ human erythrocytes²⁷. Synchrony was maintained by sorbitol treatment. Trophozoites were collected by saponin lysis, washed twice in PBS, pH 7.0, and stored at -70 °C. Trophozoite pellets were resuspended in PBS, pH 7.0, with 1.5 mM MgCl₂, 1 mM PMSF, 1 mM 1,10-phenanthroline, 0.1 mM leupeptin and 50 µM pepstatin, and mechanically disrupted with a Dounce homogenizer. A red-cell extract was prepared by homogenizing 100 µl washed, packed A⁺ human red cells in 10 ml PBS as above. Stock solutions of 2 mM haematin in 0.01 M NaOH were prepared fresh; 3 ml of this stock was added to 1 µCi ¹⁴C-haematin (Leeds Radiopharm, UK) for the experiments in *b*. At the end of each experiment free haem was removed by extracting the sample in 1 ml of 2% SDS in 0.1 M sodium bicarbonate, pH 9.1. Haemozoin was pelleted by centrifugation, washed in bicarbonate buffer, and incubated overnight in 1 mg Protease E (Sigma) in 50 mM Tris-HCl, pH 7.5. Haemozoin was again recovered by centrifugation, and the amount of haem incorporated determined either by the pyridine haemochrome method²⁸ (*a*) or by scintillation counting (*b*).

10-fold excess of either synthetic haemozoin or *P. falciparum* haemozoin purified of any protein or lipid contaminants (see ref. 12 for experimental methods), there was no activity (data not shown). A component of the freshly prepared trophozoite pellet that is distinct from the haemozoin itself is therefore responsible for the haem polymerase activity.

The pellet fraction obtained from homogenized trophozoites was used to further characterize the haem polymerase activity. Product formation was linear with both time (Fig. 3a) and the amount of protein added (Fig. 3b). The activity increased slightly as the ionic strength of the buffer was raised (data not shown), although considerable product formation occurred in all samples (50 mM–1.0 M sodium acetate, pH 5.0). The pH optimum for the activity was found between pH 5.0 and 6.0 (Fig. 3c), which covers the estimated pH of malaria food vacuoles⁴. As the pH was raised above 6.0 the activity fell dramatically, possibly because of ionization of the carboxylate side chains of the haem substrate. The linear time and protein dependence of haemozoin formation are consistent with the presence of a novel protease-resistant haem polymerase enzyme in malaria trophozoites, which tightly associates with the haemozoin product during isolation. Several haemozoin-associated proteins have been described previously^{15,16}, and it is likely that one of these is haem polymerase.

Quinoline-containing drugs accumulate in intracellular acid vesicles because of their weak base properties^{3–5,17}. Many members of this family (including chloroquine, quinine and amodiaquin) have significant antimalarial activity, although there is no satisfactory explanation for their specific toxicity. It has been suggested that they may act by raising intravacuolar pH^{4,5,17,18}, although it is disputed whether this occurs at relevant drug concentrations¹⁹. This theory also fails to explain the relative inactivity of various stereoisomers and congeners of chloroquine and quinine that retain the weak base property. Another suggestion has been that chloroquine forms a toxic complex with haem inside the food vacuole²⁰, although it has not been possible to show that sufficient free haem is available²¹. Significant chloroquine toxicity is limited to the intraerythrocytic phase of the parasite, and is most pronounced when the ring/trophozoite stage parasites are actively digesting haemoglobin²². This suggests that these drugs are inhibiting a crucial metabolic step specific to the parasite at this stage in its life cycle.

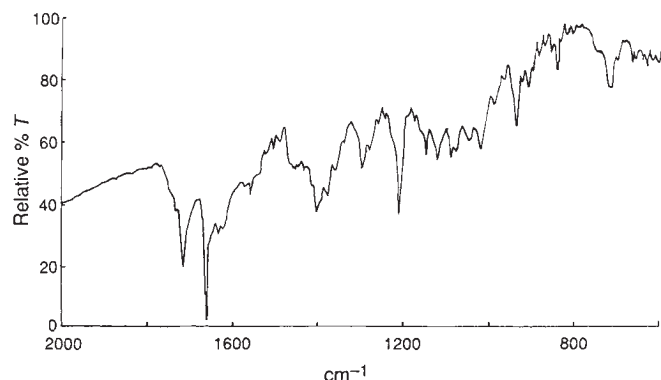


FIG. 2 Fourier-transform infrared spectrum of the haemozoin product after an enzyme assay.

METHODS. Haem polymerase in a trophozoite extract (2.3 mg protein, 47 nmol haemozoin-associated haem) was assayed with haem as described in Fig. 1a. The product remaining after SDS-bicarbonate and protease treatment was washed three times in H₂O and lyophilized. A KBr pellet was prepared, and the spectrum acquired for 30 cycles in a Fourier-transform infrared spectrometer (Perkin Elmer, model 1980). A control spectrum obtained from a trophozoite extract alone was very similar to that shown, except that the peaks were less intense.

Chloroquine inhibits the haem polymerase activity present in an extract of *P. falciparum* trophozoites (50% inhibitory concentration (IC₅₀) is 120 μ M; Fig. 4a). This inhibition occurs at a pH and drug concentration similar to that estimated in the malaria food vacuole. In similar experiments, 3-methyl chloroquine (an antimalarially active congener of chloroquine) gave 60% inhibition of haem polymerase at 1 mM, whereas 8-chloroquine (inactive as an antimalarial) gave no detectable inhibition at 5 mM (data not shown). Chloroquine also inhibited haem polymerase when a haemoglobin-rich red cell lysate was

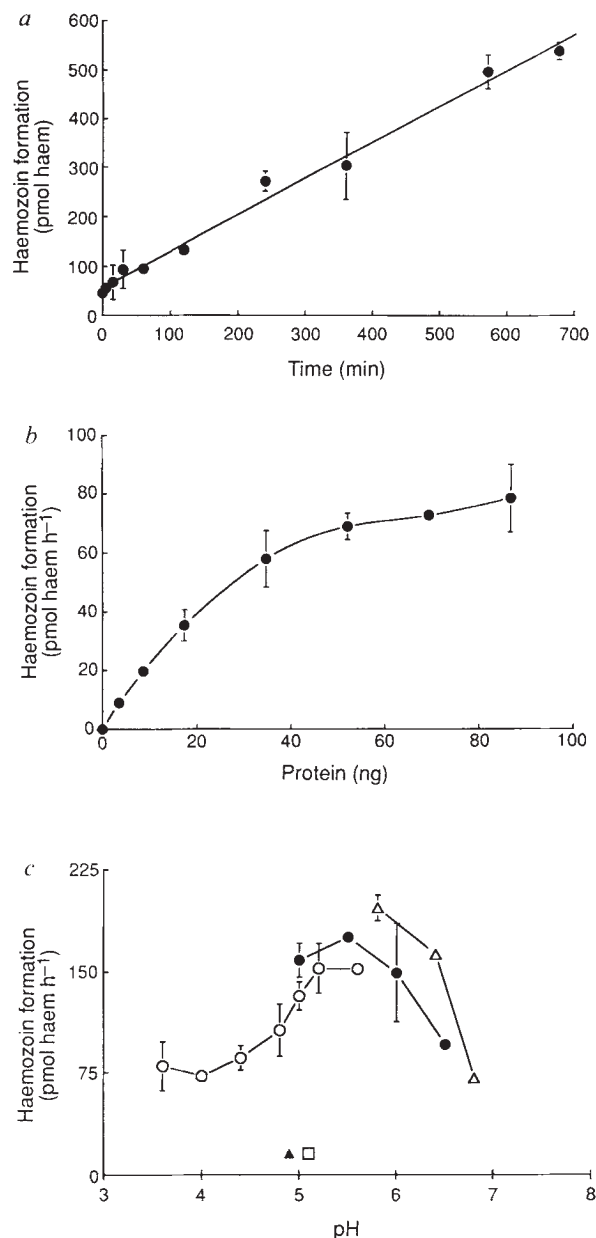


FIG. 3 Initial characterization of haem polymerase. *a*, Time-dependent enzyme activity. *b*, Concentration-dependent enzyme activity. *c*, pH-dependent enzyme activity.

METHODS. Trophozoite extracts were prepared as described in Fig. 1, and centrifuged at 25,000g for 30 min at 4 °C. The supernatant was discarded, and the membrane-haemozoin pellet resuspended in PBS, pH 7.0. *a*, *b*, Haem polymerase activity was assayed at 500 mM sodium acetate, pH 5.0, whereas for the pH curve in *c*, 200 mM sodium acetate (○), citrate-phosphate (●) or sodium phosphate (△) were used as buffers (control incubations in the absence of enzyme are also shown: acetate (▲) and citrate (□)). The enzyme was incubated with ¹⁴C-haem for 0–12 h (*a*) or 9 h (*b*, *c*). Each data point shows mean $\pm$ s.e.m. for triplicate assays.

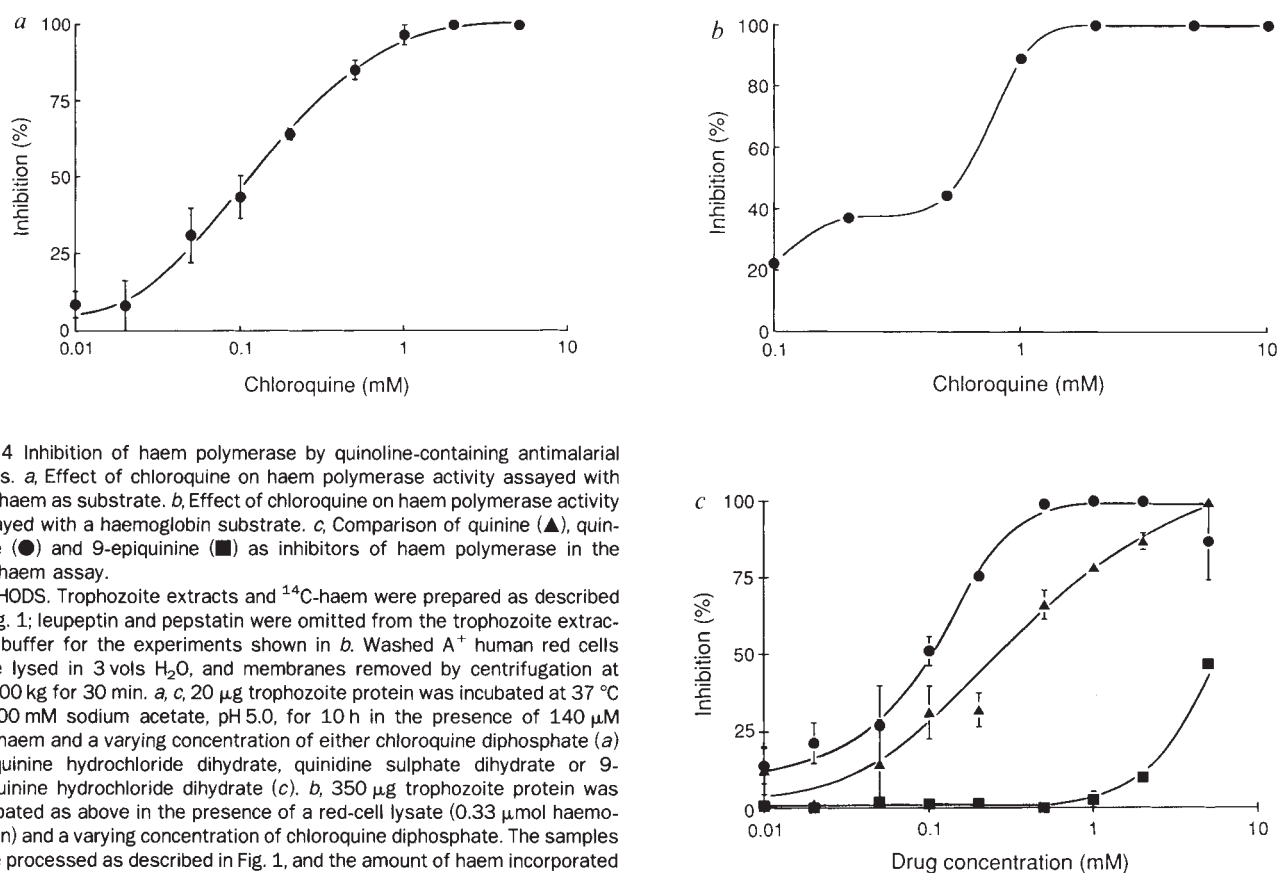


FIG. 4 Inhibition of haem polymerase by quinoline-containing antimalarial drugs. *a*, Effect of chloroquine on haem polymerase activity assayed with ^{14}C -haem as substrate. *b*, Effect of chloroquine on haem polymerase activity assayed with a haemoglobin substrate. *c*, Comparison of quinine (▲), quinidine (●) and 9-epiquinine (■) as inhibitors of haem polymerase in the ^{14}C -haem assay.

METHODS. Trophozoite extracts and ^{14}C -haem were prepared as described in Fig. 1; leupeptin and pepstatin were omitted from the trophozoite extraction buffer for the experiments shown in *b*. Washed A⁺ human red cells were lysed in 3 vols H₂O, and membranes removed by centrifugation at 25,000 kg for 30 min. *a*, *c*, 20 μg trophozoite protein was incubated at 37 °C in 500 mM sodium acetate, pH 5.0, for 10 h in the presence of 140 μM ^{14}C -haem and a varying concentration of either chloroquine diphosphate (*a*) or quinine hydrochloride dihydrate, quinidine sulphate dihydrate or 9-epiquinine hydrochloride dihydrate (*c*). *b*, 350 μg trophozoite protein was incubated as above in the presence of a red-cell lysate (0.33 μmol haemoglobin) and a varying concentration of chloroquine diphosphate. The samples were processed as described in Fig. 1, and the amount of haem incorporated determined by scintillation counting (*a*, *c*) or the pyridine haemochrome method (*b*). Results are expressed as per cent inhibition relative to haemozoin formation in a drug-free control, and show either mean $\pm$ s.e.m. for triplicates (*a*, *c*) or single assays (*b*).

used as the source of haem for the enzyme (Fig. 4*b*). The quinoline antimalarial amodiaquin, which unlike chloroquine does not form a complex with free haem²³, also inhibits haem polymerase (IC_{50} , 250 μM ; data not shown). This suggests that the enzyme inhibition cannot be explained by a substrate-drug interaction, although further experiments with the more active monodesethyl metabolite of amodiaquin are required to confirm this.

Although the weak base properties of quinine and its isomers, quinidine and epiquinine, are identical, their antimalarial activities vary by several hundredfold²⁴. These drugs also varied in their capacity to inhibit malaria haem polymerase (Fig. 4*c*); quinidine (IC_{50} , 90 μM) was the best inhibitor, followed by quinine (IC_{50} , 300 μM) and epiquinine (IC_{50} > 5 mM). This is the same rank order as their antimalarial activity^{24,25}, consistent

with the hypothesis that haem polymerase is the physiological target for these drugs.

It is estimated that during chemotherapy chloroquine reaches millimolar concentrations in the food vacuole of a susceptible malaria trophozoite²¹. We propose that this will inhibit haem polymerase, thereby disrupting the ordered conversion of haemoglobin-bound haem into haemozoin. Haem is highly toxic for malaria proteases^{9,26}, and it would therefore be expected to rapidly block further haemoglobin degradation. This could explain the specific vulnerability of growing intraerythrocytic malaria parasites to quinoline-containing drugs. A detailed understanding of the mechanism of action of haem polymerase would open the possibility of rational design of new classes of antimalarial agents. □

Received 18 September; accepted 22 October 1991.

- WHO Technical Report Series 805 *Practical Chemotherapy of Malaria* (WHO, Geneva, 1990).
- Peters, W. *Chemotherapy and Drug Resistance in Malaria* (Academic, London, 1987).
- Yayon, A., Cabatnick, Z. & Ginsburg, H. *EMBO J.* **3**, 2695–2700 (1984).
- Krogstad, D. J., Schlesinger, P. H. & Gluzman, I. Y. *J. Cell Biol.* **101**, 2302–2309 (1985).
- Krogstad, D. J. & Schlesinger, P. H. *New Engl. J. Med.* **317**, 542–549 (1987).
- Yayon, A., Timberg, R., Friedman, S. & Ginsburg, H. *J. Protozool.* **31**, 367–370 (1984).
- Goldberg, D. E., Slater, A. F. G., Cerami, A. & Henderson, G. B. *Proc. natn. Acad. Sci. U.S.A.* **87**, 2931–2935 (1990).
- Tappel, A. L. *Archs Biochem. Biophys.* **44**, 378–395 (1953).
- Vander Jagt, D. L., Hunsaker, L. A. & Campos, N. M. *Molec. Biochem. Parasitol.* **18**, 389–400 (1986).
- Yasuhara, T., Mori, M., Walamatsu, K. & Kubo, K. *Biochem. biophys. Res. Commun.* **178**, 95–103 (1991).
- Scheibel, L. W. & Sherman, I. W. in *Malaria: Principles and Practice of Malariology* (eds Wernsdorfer, W. H. & McGregor, I.) 219–252 (Churchill-Livingstone, Edinburgh, 1988).
- Slater, A. F. G. et al. *Proc. natn. Acad. Sci. U.S.A.* **88**, 325–329 (1991).
- Lemberg, R. & Legge, J. W. *Hematin Compounds and Bile Pigments* (Interscience, New York, 1949).
- Fitch, C. D. & Kanjanangulpan, P. *J. biol. Chem.* **262**, 15552–15555 (1987).
- Yamada, K. A. & Sherman, I. W. *Exp. Parasitol.* **48**, 61–74 (1979).

- Ashong, J. O., Blench, I. P. & Warhurst, D. C. *Trans. R. Soc. trop. Med. Hyg.* **83**, 167–172 (1989).
- Homewood, C. A., Warhurst, D. C., Peters, W. & Baggeley, V. C. *Nature* **235**, 50–52 (1972).
- Krogstad, D. J. & Schlesinger, P. H. *Biochem. Pharmac.* **35**, 547–552 (1986).
- Ginsburg, H., Nissani, E. & Krugliak, M. *Biochem. Pharmac.* **38**, 2645–2654 (1989).
- Fitch, C. D., Chevli, R., Phillips, G., Pfaller, M. A. & Krogstad, D. J. *Antimicrob. Agents Chemother.* **21**, 819–822 (1982).
- Ginsburg, H. & Geary, T. G. *Biochem. Pharmac.* **36**, 1567–1576 (1987).
- Zhang, Y., Asante, K. S. O. & Jung, A. J. *Parasit.* **72**, 830–836 (1986).
- Warhurst, D. C. *Ann. trop. Med. Parasitol.* **81**, 65–67 (1987).
- Warhurst, D. C. *Biochem. Pharmac.* **30**, 3323–3327 (1981).
- White, N. J. et al. *Lancet* **ii**, 1069–1071 (1981).
- Vander Jagt, D. L., Hunsaker, L. A. & Campos, N. M. *Biochem. Pharmac.* **36**, 3285–3291 (1987).
- Trager, W. & Jensen, J. B. *Science* **193**, 673–675 (1976).
- Fuhrhop, J. H. & Smith, K. M. in *Porphyrins and Metalloporphyrins* (ed. Smith, K. M.) 757–889 (Elsevier, Amsterdam, 1975).

ACKNOWLEDGEMENTS. We thank D. Goldberg and N. Le Trang for encouragement, O. Bloom for technical assistance, J. Williams and W. Trager for the parasite clone, D. Cowburn for access to his IR spectrometer, K. Barry Sharpless (MIT) for the epiquinine and B. Ellis (Walter Reed) for the chloroquine congeners. This work was supported by the Rockefeller Foundation and the NIH.

Extrathymic positive selection of $\alpha\beta$ T-cell precursors in nude mice

Daniel E. Speiser, Ursula Stübi & Rolf M. Zinkernagel

Laboratory for Experimental Pathology, Institute of Pathology, University Hospital, CH-8091 Zürich, Switzerland

T LYMPHOCYTES expressing $\alpha\beta$ T-cell receptors with sufficient affinity to major histocompatibility complex (MHC) molecules expressed on thymus epithelial cells are positively selected and mature to functional T cells¹⁻⁶. But several studies have demonstrated that athymic nude mice⁷⁻⁹ grafted with MHC-incompatible thymuses developed T cells specific for nude host rather than thymic MHC¹⁰⁻¹⁴. We examined this paradox by analysing the specificity of T lymphocytes derived from nude mice. We report here that nude T lymphocyte precursors transferred to allogeneic SCID (severe combined immunodeficiency) mice with a functioning thymus (but lacking T or B cells¹⁵⁻¹⁷) generated host MHC-restricted effector T cells but also contained T cells restricted to donor MHC. If nude T cells were depleted from nude lymphohaemopoietic donor cells before or after transfer, only host MHC-specific T cells matured. The results may explain the unusual MHC specificities of nude T lymphocytes described in earlier studies¹⁰⁻¹⁴ and demonstrate two separate differentiation steps: in nude mice, T cells may be positively selected for self-MHC restriction specificity extrathymically; then a functional thymus is required for efficient T cell maturation.

Nude mice contain many T cells¹⁸⁻²⁴ but cannot generate pathogen-specific T-cell responses because they have no thymus⁷⁻⁹. SCID mice lack functional lymphocytes owing to their inability to rearrange and express T-cell receptors (TCR) and immunoglobulin genes¹⁵⁻¹⁷. To complement the two distinct defects, bone marrow chimaeras were made by intravenous injection of bone marrow cells from athymic *nu/nu* mice into untreated C.B-17/Icr *scid/scid* (SCID, H-2^d) mice²⁵. They were tested 12-20 weeks later for the generation of virus-specific cytotoxic T lymphocytes (CTL). SCID mice reconstituted with H-2-identical bone marrow cells from nude BALB/c mice successfully generated H-2^d-restricted lymphocytic choriomeningitis virus (LCMV)-specific CTL (Table 1a).

We have recently demonstrated that SCID mouse chimaeras reconstituted with T-cell depleted bone marrow from allogeneic nude C57BL/6 (*nu/nu* B6, H-2^b) mice generated CTL restricted to H-2^d of the SCID host²⁵, confirming that the host's thymus directs T-cell MHC restriction specificity¹⁻⁶. But when SCID mice were reconstituted with unpurified allogeneic H-2^b nude bone marrow plus spleen cells they generated host-restricted but surprisingly also nude donor-restricted LCMV-specific CTL. Therefore does the ability of allogeneic SCID chimaeras to

mount donor H-2^b-restricted CTL correlate with their reconstitution with nude T cells? To test this we directly compared CTL responses of SCID mice reconstituted with either T-cell depleted or alternatively with nondepleted nude bone marrow plus spleen cells. Whereas the former generated CTL lysing LCMV-infected D2 (H-2^d) fibroblast targets only, the latter also developed CTL that lysed LCMV-infected MC57G (H-2^b) fibroblasts (Table 1b). LCMV-infected H-2^b-expressing L929 target fibroblasts were not lysed. Thus, SCID mice with T-cell-depleted allogeneic donor cells lacked donor MHC-restricted CTL, whereas SCID mice with unpurified donor cells generated strong H-2^d but also H-2^b-specific cytotoxicity comparable to (B6 $\times$ BALB/c)_{F1} H-2^{bxd} control mice.

The result that nude T cells were preprogrammed to express nude H-2^b restriction specificity was supported by *nu/nu* B6 $\rightarrow$ SCID chimaeras which were T-cell depleted *in vivo* 20 weeks after reconstitution by treatment with cytotoxic anti-CD4 plus anti-CD8 monoclonal antibodies. Eight weeks later they showed SCID H-2^d but not nude H-2^b-restricted CTL (Table 1c). This finding also excluded the possibility that nude non-T cells migrating to the SCID host thymus were able to select H-2^b-restricted T cells or that allogeneic stimulation may have caused 'abnormal'²⁶ induction of maturation¹⁰, a possibility which was unlikely because of the immunoincompetence of SCID mice. Thus, in the allogeneically reconstituted SCID mice, nude T-lymphocyte precursors gave rise to nude donor H-2^b-restricted cytotoxic T cells. Finally, thymectomized chimaeric SCID mice generated neither host H-2^d- nor donor H-2^b-restricted CTL activity, demonstrating that nude T-cell precursors required a functional thymus for full differentiation (Table 1c).

Crossreactive T cells with restriction specificity both for H-2^d and for H-2^b may have been responsible for the observed nude donor H-2^b-restricted cytotoxicity. But cold target competition experiments revealed that the H-2^b-restricted killing could be blocked with unlabelled infected MC57G (H-2^b) but not with unlabelled infected D2 (H-2^d) fibroblasts (Fig. 1). Also, experiments with target cells coated with synthetic viral peptides (Table 1) showed that the H-2^b-restricted cytotoxicity was specific for the LCMV glycoprotein-derived peptides, whereas the H-2^d-restricted killing was specific for the peptide epitope from the LCMV nucleoprotein^{27,28}. Thus, two distinct T-cell populations were present, one with restriction specificity to H-2^d and the other to H-2^b.

The protective antiviral capacity of nude-derived T cells was determined by measurements of virus titres in spleens 15 days after LCMV infection (Table 2). The SCID mice with non-depleted but not those with T-cell-depleted *nu/nu* B6 donor cells were able to eliminate LCMV. The latter developed a virus carrier state because after 60 days the virus titres remained high (data not shown). As LCMV clearance is essentially CTL-dependent²⁹, the data suggest that nude donor H-2^b-restricted T cells were required to eliminate LCMV from infected donor H-2^b cells.

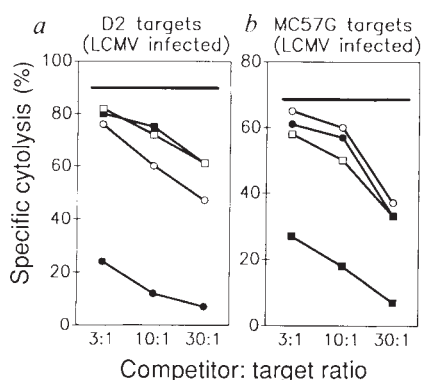


FIG. 1 Cold target inhibition of CTL activity from lymph node cells of a SCID (H-2^d) mouse reconstituted with nondepleted bone marrow plus spleen cells from *nu/nu* B6 (H-2^b) mice. 25 weeks after priming *in vivo* with 100 p.f.u. LCMV, 3×10^6 effector spleen cells were restimulated *in vitro* for 5 days with LCMV-infected BALB/c (a) or B6 (b) macrophages (3×10^5). The ⁵¹Cr-release assay was done in duplicate cultures, each with 2×10^5 effectors and 10^4 target cells. Competitor cells were titrated into the wells as indicated. Spontaneous release (4h) was 18% for D2 cells and 22% for MC57G cells. Horizontal line, specific cytotoxicity (%) in the absence of competitor. Cytotoxicity of uninfected control targets was less than 8%. Cold target competitors: MC57G-LCMV (■); MC57G (□); D2-LCMV (●); D2 (○).

TABLE 1 T-cell depletion before or after transfer of allogeneic *nu/nu* B6 lymphoid donor cells abrogates donor H-2^b-restricted cytotoxic T-cell responses

Responder spleen cells from controls or chimaeras			Specific lysis of targets (%)					
(Donor → SCID H-2 ^d host)	T-cell depletion		MC57G (H-2 ^b)		D2 (H-2 ^d)		L929 (H-2 ^k)	YAC-1
	<i>in vitro</i>	<i>in vivo</i>	LCMV peptides	Normal	LCMV peptide	Normal	LCMV infected	Normal
(a) (B6×BALB/c)F ₁			47	4	94	<1	4	7
			33	<1	89	<1	2	3
			12	1	69	<1	<1	3
			1	<1	<1	<1	11	23
			<1	<1	3	1	8	23
			<1	<1	2	1	3	12
	SCID		<1	<1	<1	5	n.d.	28
			<1	<1	<1	4		16
			<1	<1	<1	<1		3
			<1	<1	96	<1	6	8
			<1	<1	87	<1	4	4
			3	<1	55	<1	<1	<1
(b) <i>nu/nu</i> B6→SCID	+	-	<1	<1	75	1	1	<1
			<1	<1	68	<1	7	<1
			<1	<1	41	<1	3	<1
	+	-	3	1	100	6	9	22
			<1	4	86	2	4	19
			6	5	46	3	6	8
	-	-	62	<1	83	1	1	<1
			31	<1	68	2	3	<1
			24	<1	40	2	3	<1
	-	-	53	6	83	2	8	<1
			39	2	82	2	<1	<1
			32	6	39	<1	<1	<1
(c) <i>nu/nu</i> B6→SCID	-	+	<1	<1	96	<1	8	<1
			<1	<1	83	<1	4	<1
			<1	<1	53	<1	7	<1
	-	-	<1	<1	<1	2	n.d.	27
			<1	<1	<1	<1		15
			<1	3	<1	<1		14

LCMV-specific cytotoxic T-cell responses of SCID (H-2^d) mice reconstituted intravenously with bone marrow and spleen cells from nude BALB/c (*nu/nu* BALB/c, H-2^d) or nude C57BL/6 (*nu/nu* B6, H-2^b) mice. In general, the data are from SCID mice 12–20 weeks after reconstitution. T-cell depletion was done either *in vitro* before transfer of *nu/nu* donor cells or *in vivo* after transfer but 8 weeks before the test. *In vitro* cytotoxicity was assessed 9 days after intravenous infection with 100 plaque-forming units (p.f.u.) of LCMV in standard ⁵¹Cr-release assays^{10,25} at effector to target ratios of 70/23/8. MC57G, D2, L929 or NK-sensitive YAC-1 target cells were uncoated ('normal'), or specifically coated ('LCMV peptide(s)') at 37 °C for 90 min with 50 μM synthetic viral peptide epitopes (MC57G cells with the two LCMV glycoprotein-derived peptides amino acids 32–42 and 275–286 (ref. 27); D2 cells with the LCMV nucleoprotein-derived peptide amino acids 118–126 (ref. 28)). LCMV-infected target cells revealed similar data as targets coated with the corresponding viral peptide(s) and therefore only the latter results are shown. L929 LCMV-infected cells were specifically lysed by LCMV-immune spleen cells from B10.BR mice (54/32/24% specific lysis). Spontaneous release of target cells (5 h incubation) was <25%. Data from each responder shown are representative for 4–6 mice or chimaeras per group. LCMV-specific cytotoxicity was also investigated under limiting dilution conditions: in spleen cells, CTL frequencies determined on infected MC57G/D2 cells were $4 \times 10^{-4}/8 \times 10^{-3}$ for nondepleted *nu/nu* B6→SCID, $<10^{-5}/9 \times 10^{-3}$ for *in vitro* T-cell depleted *nu/nu* B6→SCID, $4 \times 10^{-2}/<10^{-5}$ for B6, $<10^{-5}/2 \times 10^{-3}$ for BALB/c, and $<10^{-5}/<10^{-5}$ for *nu/nu* B6 mice. Flow cytometry of lymph node and spleen cells from the allogeneically reconstituted SCID mice revealed that >99% of the T and B cells expressed H-2^b but not H-2^d, whereas about 50% of the nonlymphocytes were of H-2^b and 50% of H-2^d haplotype. Furthermore, the relative percentages of cells expressing IgM, CD3, CD4, CD8, Vβ6 and Vβ11 were comparable to normal B6 mice and no significant differences between T-cell depleted and nondepleted chimaeras were detectable. Also, the SCID chimaeras did not show increased percentages of γδTcR⁺ or Vβ8⁺ T cells, both of which may be overexpressed in nude mice^{20,23,24}, n.d., Not done. Methods. BALB/c (H-2^d), C57BL/6 (B6, H-2^b) and B10.BR (H-2^k) mice were purchased from the Institut für Zuchtthygiene, Tierspital, University of Zürich, Switzerland. Nude BALB/c and nude C57BL/6 were obtained from Bolmholtegaard Ltd, Ry, Denmark. C.B-17/lcr *scid/scid* (SCID, H-2^d) mice (unable to rearrange and express TCR and immunoglobulin genes^{16,17}) were obtained from the universities of Ulm and Munich, Germany, bred locally under standard pathogen-free conditions and screened for absence of immunoglobulins in sera. They were reconstituted at the age of 8–12 weeks with $5\text{--}20 \times 10^6$ donor *nu/nu* bone marrow plus *nu/nu* spleen cells which were mixed (ratio 1:1) and where indicated, *in vitro* T-cell depleted twice using Thy-1.2-specific monoclonal antibodies HO 13.4-9 and AT-83, and a polyclonal mouse anti-thymocyte serum¹ plus rabbit complement resulting in <0.1% CD3⁺ cells. In the cases of 'nondepletion', donor cells were treated similarly but without Thy-1.2-specific antibodies or sera and contained about 5% CD3⁺ cells. In each reconstitution at least 30% of nude lymphohaemopoietic cells were derived from nude donor mice older than 18 weeks, the remaining being 10–18 weeks. *In vivo* T-cell depletion of SCID and control mice was achieved using the cytotoxic antibodies YTS 191.1 (CD4-specific) plus YTS 169.4 (CD8-specific) injected intravenously twice at an interval of 2 days with 1 mg per antibody and mouse. Chimaerism was assessed by double fluorescence-analysis with H-2 haplotype-specific and CD4- or CD8-specific antibodies.

To assess the restriction specificity of nude T-helper cells, *nu/nu* B6→SCID chimaeras were tested for neutralizing antibody responses to the vesicular stomatitis virus (VSV) *in vivo*. All chimaeras generated T helper cell-independent VSV-specific IgM antibodies 4 days after infection with 5×10^7 plaque-

forming units (Table 3). But the strictly T helper-dependent³⁰ IgG responses were absent in chimaeras with T-cell depleted allogeneic *nu/nu* B6 donor cells, indicating inefficient *in vivo* T-B cell cooperation, presumably owing to the mismatch of T-cell restriction specificity (to H-2^d) and MHC products

TABLE 2 T-cell depletion from allogeneic *nu/nu* B6 lymphoid donor cells impairs CTL-driven virus elimination

Control mice or chimaeras		Virus titre (log ₁₀ p.f.u. per g spleen)
(Donor → SCID H-2 ^d host)	T-cell depletion <i>in vitro</i>	
BALB/c		<2
B6		<2
SCID		6.9±0.1
<i>nu/nu</i> B6 → SCID	+	6.2±0.5
<i>nu/nu</i> B6 → SCID	—	<2

Virus titres in spleens from controls or SCID chimaeras (mean ± s.e.m. of 3 or 4 mice per group) were measured 15 days after LCMV infection intravenously with 2×10^6 p.f.u. LCMV-WE in a standard plaque assay²⁸. '<2', the titres were below the detection level of 2 log₁₀. T-cell depletion and SCID mouse reconstitution are described in the legend to Table 1.

expressed by B cells (H-2^b)²⁵. Furthermore, thymectomized SCID chimaeras failed to generate VSV-specific IgG antibodies, confirming that preprogrammed nude T cells required a thymus for full maturation. In conclusion, the nude H-2^b-restricted T-helper cells cooperated with donor H-2^b B cells, in contrast to the newly developed H-2^d-restricted T cells.

There are many T-cell-marker positive cells in peripheral lymphatic tissues of nude mice¹⁸⁻²⁴. Although they contain alloreactive and trinitrophenyl-specific cytotoxic T-cell precursors³¹⁻³⁴ as well as T-helper activity for antibody production³⁵, they do not generate protective MHC-restricted pathogen-specific T cells⁷⁻⁹. How can nude mice select and propagate precursor T cells? Several possibilities have been excluded, such as placental passage of maternal lymphocytes¹⁹, or that the thymic rudiments may promote T-cell maturation¹⁰. Other tissues such as gut-associated epithelial cells^{36,37} may possibly

be involved. Furthermore, T lymphocytes with affinity to self MHC may survive preferentially and accumulate. In any case, the existence of preprogrammed T-cell populations may be relevant and of possible therapeutic relevance in human bone marrow transplantation^{38,39}.

Our data may explain the paradoxical findings that T-cell restriction specificity of nude mice engrafted with an allogeneic fetal or newborn thymus preferentially, if not exclusively, corresponded to the MHC haplotype of the nude host¹⁰⁻¹⁴. These results did not fit those from (H-2^b × H-2^d)F₁ *nu/nu* mice reconstituted with one parental thymus graft which exhibited virtually exclusively parental thymic H-2 restriction specificity^{40,41}. Accordingly, the extremely selective conditions of *nu/nu* B6 mice reconstituted with H-2^d thymic grafts favouring H-2^b-specific interactions and expansions (all antigen presentation was with H-2^b) enhanced the salvage pathway for the nude H-2-selected but not fully matured T cells. In F₁ nude mice the obviously much more efficient direct thymic pathway was dominant.

The data confirm that after positive selection, lymphocytes must undergo further thymus-dependent differentiation^{13,42-45}, including clonal deletion²¹⁻²³. Furthermore, thymic hormones may be necessary in promoting nude precursor T-cell differentiation^{12,46}. Our experiments reveal separable stages of T-cell maturation. Positive selection of self-MHC restriction happens most efficiently in the thymus, but may also occur extrathymically as shown here for nude T cells. However, further maturation seems much more efficient in the presence of a thymus. □

Received 22 July; accepted 3 October 1991.

- Zinkernagel, R. M. *et al.* *J. exp. Med.* **147**, 882-896 (1978).
- Bevan, M. J. & Fink, P. J. *Immun. Rev.* **42**, 4-19 (1978).
- Spren, J., Lo, D., Gao, K. E. & Ron, Y. *Immun. Rev.* **101**, 173-190 (1988).
- Schwartz, R. H. *Cell* **57**, 1073-1081 (1989).
- Kisielow, P., Teh, H. S., Blüthmann, H. & von Boehmer, H. *Nature* **335**, 730-733 (1988).
- Sha, W. C. *et al.* *Nature* **335**, 271-274 (1988).
- Pantelouris, E. M. *Immunology* **20**, 247 (1971).
- Wortis, H. H. *Contemp. Top. Immunobiol.* **3**, 243 (1974).
- Kindred, B. *Prog. Allergy* **26**, 137-238 (1979).
- Zinkernagel, R. M. *et al.* *J. exp. Med.* **151**, 376-399 (1980).
- Lake, J. P., Andrew, M. E., Pierce, C. W. & Braciale, T. J. *J. exp. Med.* **152**, 1805-1810 (1980).
- Kruisbeek, A. M., Sharrow, S. O., Mathieson, B. J. & Singer, A. *J. Immun.* **127**, 2168-2176 (1981).
- Kast, W. M., Waal, L. P. de & Melief, C. J. M. *J. exp. Med.* **160**, 1752-1763 (1984).
- Longo, D. L., Kruisbeek, A. M., Davis, M. L. & Matis, L. A. *Proc. natn. Acad. Sci. U.S.A.* **82**, 5900-5904 (1985).
- Bosma, G. C., Custer, R. P. & Bosma, M. J. *Nature* **301**, 527-530 (1983).
- Schuler, W. *et al.* *Cell* **46**, 963-972 (1986).
- Malynn, B. A. *et al.* *Cell* **54**, 453-460 (1988).
- Raff, M. C. *Nature* **246**, 350-352 (1973).
- Loor, F., Amstutz, H., Hägg, L. B., Mayor, K. S. & Roelants, G. E. *Eur. J. Immun.* **6**, 663-666 (1976).
- MacDonald, H. R., Lees, R. K., Bron, C., Sordat, B. & Miescher, G. *J. exp. Med.* **166**, 195-209 (1987).
- Hodes, R. J., Sharrow, S. O. & Solomon, A. *Science* **246**, 1041-1044 (1989).
- Fry, A. M., Jones, L. A., Kruisbeek, A. M. & Matis, L. A. *Science* **246**, 1044-1046 (1989).
- Rocha, B. *Eur. J. Immun.* **20**, 919-925 (1990).
- Lake, J. P., Pierce, C. W. & Kennedy, J. D. *Cell. Immun.* **135**, 259-265 (1991).
- Zinkernagel, R. M., Rüdel, E., Althage, A., Hengartner, H. & Reimann, G. *J. exp. Med.* **168**, 1187-1192 (1988).
- Hamaoka, T., Osborne, D. P. & Katz, D. H. *J. exp. Med.* **137**, 1393-1404 (1973).
- Oldstone, M. B. A., Whitton, J. L., Lewicki, H. & Tishon, A. *J. exp. Med.* **168**, 559-570 (1988).
- Schulz, M. *et al.* *Eur. J. Immun.* **19**, 1657-1667 (1989).
- Buchmeier, M. J., Welsh, R. M., Dutko, F. J. & Oldstone, M. B. A. *Adv. Immun.* **30**, 275-312 (1980).
- Charan, S. & Zinkernagel, R. M. *J. Immun.* **136**, 3057-3061 (1986).
- Gillis, S., Union, N. A., Baker, P. E. & Smith, K. A. *J. exp. Med.* **149**, 1460-1476 (1979).
- Maryanski, J. L., MacDonald, H. R., Sordat, B. & Cerottini, J.-C. *J. Immun.* **126**, 871-876 (1981).
- Hünig, T. & Bevan, M. J. *J. exp. Med.* **152**, 688-702 (1980).
- Ando, I. & Hurme, M. *Nature* **289**, 494-495 (1981).
- Ishikawa, H. & Saito, K. *J. exp. Med.* **151**, 965-968 (1980).
- Klein, J. R. *J. exp. Med.* **164**, 309-314 (1986).
- Guy-Grand, D. *et al.* *J. exp. Med.* **173**, 471-481 (1991).
- Keever, C. A. *et al.* *Blood* **73**, 1340-1350 (1989).
- Soiffer, R. J. *et al.* *Blood* **75**, 2076-2084 (1990).
- Hedrick, S. M. & Watson, J. *J. exp. Med.* **150**, 646-652 (1979).
- Zinkernagel, R. M., Althage, A. & Callahan, G. *J. exp. Med.* **150**, 693-697 (1979).
- Ohashi, P. S., Pircher, H. P., Bürki, K., Zinkernagel, R. M. & Hengartner, H. *Nature* **346**, 861-863 (1990).
- Guidos, C. J., Danska, J. S., Fathman, C. G. & Weissman, I. L. *J. exp. Med.* **172**, 835-845 (1990).
- Shortman, K., Vremec, D. & Egerton, M. *J. exp. Med.* **173**, 323-332 (1991).
- Boyd, R. L. & Hugo, P. *Immun. Today* **12**, 71-79 (1991).
- Stutman, O. *Immun. Rev.* **42**, 138-184 (1978).

ACKNOWLEDGEMENTS. CB-17/*lcr scid/scid* mice were provided by H. Wagner and A. Kleinschmidt. We thank H. Hengartner, U. Kalinke, T. Kündig, S. Oehen, P. Ohashi and H. P. Pircher for discussion, A. Althage for technical help and P. Autenried, T. Rülke, A. Mathis and J. Gübeli for maintenance of the SCID colony. This work was supported in part by the Swiss NSF and the Radium Stiftung Zürich.

TABLE 3 T-cell depletion before or after transfer of allogeneic *nu/nu* B6 lymphoid donor cells abrogates T-cell dependent IgG neutralizing antibody responses

Responder controls or chimaeras		Neutralizing anti-VSV titres (−log ₂ × 40)	
(Donor → SCID H-2 ^d host)	T-cell depletion <i>in vitro</i> <i>in vivo</i>	IgM (day 4)	IgG (day 12)
BALB/c		9.0±0.0	8.7±0.3
B6		8.0±0.0	7.0±0.0
<i>nu/nu</i> B6		10.5±0.3	<1
SCID		<1	<1
<i>nu/nu</i> BALB/c → SCID	+	7.5±0.3	5.6±0.3
<i>nu/nu</i> B6 → SCID	+	9.5±0.3	<1
<i>nu/nu</i> B6 → SCID	—	9.5±0.5	4.3±0.6
<i>nu/nu</i> BALB/c → SCID	—	7.8±0.4	9.5±0.2
<i>nu/nu</i> B6 → SCID	—	6.0±0.0	<1
<i>nu/nu</i> B6 → thymectomized SCID	—	9.0±0.6	<1

T-cell-independent IgM and T-cell-dependent IgG neutralizing antibody responses to VSV. On day 0, responder controls or chimaeras were infected with 5×10^7 p.f.u. of VSV INDIANA strain²⁵, which was ultraviolet-inactivated by exposure to a germicidal lamp (Philips) for 2 min at a distance of 10 cm. Titrated serum samples were tested on VERO monolayers in a standard neutralization assay²⁵. Results for IgG represent serum samples pretreated with 0.1 M 2-mercaptoethanol³⁰ (1 h at 20 °C); all samples had been inactivated at 56 °C for 1 h. Mean ± s.e.m. of antibody titres from three to six individual mice and bleedings are given. T-cell depletion and SCID mouse reconstitution are described in the legend to Table 1.

Immuno-isolation of Sec7p-coated transport vesicles from the yeast secretory pathway

Alex Franzusoff, Estelle Lauzé & Kathryn E. Howell

Department of Cellular and Structural Biology,
University of Colorado Medical School, Box B-111,
4200 E. Ninth Avenue, Denver, Colorado 80262, USA

THE transport of proteins destined for post-endoplasmic reticulum locations in the secretory pathway is mediated by small vesicular carriers¹⁻³. Transport vesicles have been generated in cell-free assays from the yeast *Saccharomyces cerevisiae*, and mammalian systems⁴⁻¹⁴. Yeast genes encoding cytosolic components that participate in vesicular traffic were first identified from the collection of conditional-lethal *sec* (secretory) mutants¹⁵⁻¹⁷. Mutations in the yeast *SEC7* gene disrupt protein transport in the secretory pathway at the nonpermissive temperature¹⁸. The *SEC7* gene product is a phosphoprotein of relative molecular mass 230,000 that functions from the cytoplasmic aspect of intracellular membranes^{19,20}. We report that in a yeast cell-free transport assay, the introduction of antibodies to Sec7 protein (Sec7p) results in the accumulation of transport vesicles. These vesicles are retrieved

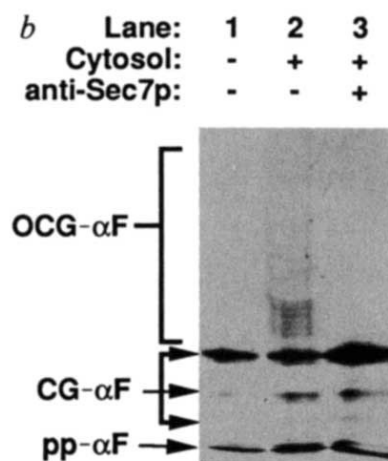
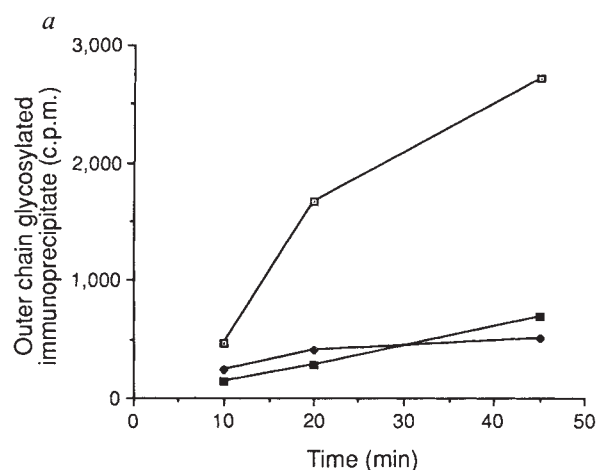
with Sec7p-specific antibodies by immuno-isolation for biochemical and electron microscopic characterization. Sec7p on the surface of the accumulated transport vesicles, in combination with previous genetic and biochemical studies^{18,20}, implicate Sec7p as part of a (non-clathrin) vesicle coat. This Sec7p-containing coat structure is proposed to be essential for vesicle budding at multiple stages in the yeast secretory pathway.

The development of yeast cell-free assays that reconstitute vesicular traffic from the endoplasmic reticulum (ER) to the early compartments of the Golgi apparatus^{21,22} provided the means to study the role of Sec7p in vesicular traffic. The transport of α -factor from the yeast ER to the Golgi apparatus *in vitro* is dependent on ATP, cytosol and elevated temperatures (Fig. 1a). The extent of transport is measured by the Golgi-dependent outer chain carbohydrate modification of the α -factor (the only radiolabelled substrate in the reaction, see legend to Fig. 1). Transport activity was blocked when affinity-purified Sec7p antibodies (Sec7-Ab), but not preimmune IgG, were added to the cell-free reactions (Fig. 1). Similar results were seen whether the cytosolic fraction was preincubated with Sec7-Ab, or if the antibodies were introduced at the transport stage of the assay.

Analysis of the labelled α -factor cargo from the cell-free reactions with or without Sec7-Ab demonstrated the accumulation of transport intermediates. A modification of the basic transport assay was introduced to fractionate the different membrane compartments in the reaction. After brief centrifugation

FIG. 1 Yeast cell-free transport assays are inhibited by Sec7p-specific antibodies. **a**, The time course for the transport reaction in the presence of wild-type cytosol plus preimmune antibodies is compared with assays carried out in the absence of cytosol, or when affinity-purified anti-Sec7p antibodies (Sec7-Ab) were added with the wild-type cytosol to the second stage of the reaction. The addition of preimmune antisera had no effect on either the rate or extent of α -factor transport. The Sec7-Ab inhibited transport efficiency to levels seen in reactions lacking cytosolic components. The time course for transport reactions with wild-type cytosol plus either preimmune or Sec7p-specific antibodies is shown. $\square$, Cytosol plus preimmune; $\blacklozenge$, cytosol plus Sec7-Ab; $\blacksquare$, no cytosol. **b**, The transport assay was done in the absence of cytosol (lane 1), in the presence of wild-type yeast cytosolic components (lane 2), and with cytosol plus Sec7p-specific antibodies (lane 3). The different forms of α -factor are indicated on the left side: the *in vitro* translated, unglycosylated precursor (pre-pro α -factor, pp- α F), the core-glycosylated α -factor polypeptides (cg- α F), and the Golgi-modified outer chain glycosylated α -factor species (ocg- α F). The addition of Sec7-Ab had no effect on the quantity of α -factor released into the MSS. But note the increased density of core-glycosylated α -factor in lane 3 compared with lanes 1 and 2.

METHODS. The assay, described previously^{21,22}, was done in two stages: translocation of α -factor into the ER, and α -factor transport from the ER to the Golgi apparatus. Radiolabelled α -factor precursors that had been translated *in vitro* with cytosolic proteins from wild-type or *sec7* yeast were translocated into the ER compartments of gently-lysed wild-type or *sec7* spheroplasts for 30 min at 11 °C. The lower temperature restricts membrane traffic from the yeast ER *in vitro*^{21,22}. After translocation, the membrane compartments were sedimented at 12,000g for 10 min, then resuspended with buffer 88 (250 mM sorbitol, 150 mM potassium acetate, 5 mM magnesium acetate, 20 mM HEPES pH 6.8) and an energy-regenerating system (40 mM creatine phosphate, 0.2 mg ml⁻¹ creatine phosphokinase, 1 mM ATP, 50 μ M GDP-mannose). Aliquots of ~60 μ g wild-type cytosolic proteins and 1–2.5 μ g antibody, either preimmune IgG or affinity-purified Sec7-Ab were added to the reaction as indicated (25 μ l total assay volume) with incubation at 20–30 °C. The Sec7-Ab was affinity purified from rabbit serum by low pH (0.1 M glycine, pH 2.5) elution of IgG from a column containing the *E. coli* produced hybrid fusion protein (TrpE–Sec7p fusion²⁰) coupled to CNBr-activated Sepharose beads. This fusion protein differed from the LacZ–Sec7p fusion protein used to immunize the rabbits²⁰. Transport of core-glycosylated α -factor to the Golgi apparatus resulted in outer-chain glycosylation, which was measured by the precipitation of radiolabelled protein with specific antisera to the outer chain carbohydrates (α -1,6 mannose antibodies) followed by scintillation counting^{18,21,22}. In **b**, after 45 min incubation at 20 °C in the (second) transport stage of the assay, the reactions were centrifuged for 45 s at 12,000g. This medium-speed supernatant (MSS) contains vesicles and Golgi compartments, but lacks the majority of ER,



which sediments with the pellet fraction. The MSS fraction was solubilized in Laemmli SDS sample buffer, and the proteins were resolved on an 11.25% SDS-gel. The autoradiogram of radiolabelled α -factor polypeptides found in the MSS fraction is displayed.

TABLE 1 Quantitative analysis of immuno-isolated vesicles

Sample	Small vesicles		Large vesicles	
	Average size (nm)	Number per field	Average size (nm)	Number per field
MSS				
No antibody	40.0 ± 10.9* (57)†	14 ± 5 (6)‡	183.0 ± 104.2* (62)†	16 ± 4 (6)‡
Sec7-antibody	42.4 ± 8.5 (73)	40 ± 18 (6)	177.0 ± 64.2 (67)	36 ± 3 (6)
Immuno-isolates				
Pre-immune antibody	48.4 ± 13.3 (9)	1 ± 1 (9)	112.7 ± 54.5 (43)	5 ± 2 (9)
Sec7-antibody	42.4 ± 12.1 (94)	28 ± 17 (10)	170.9 ± 90.9 (46)	4 ± 2 (10)

* Standard deviation of the mean.

† Number of vesicles measured, no more than 10 per electron microscope (EM) field.

‡ Number of EM fields analysed.

of the reaction mixture, transport vesicles, Golgi membranes and cytosol are recovered from the medium-speed supernatant (MSS), whereas most of the ER sediments^{11,12}. The form of the membrane-enclosed (protease-resistant) cargo that was released from the ER into the MSS was analysed by immunoprecipitation of the labelled α -factor, followed by SDS-PAGE and autoradiography. In the absence of cytosol, only a background level of core-glycosylated α -factor was observed (CG- α F; Fig. 1b, lane 1). In the presence of cytosolic proteins, a more extensively glycosylated population of α -factor was recovered from the MSS, indicating that the substrate gained access to the Golgi (OCG- α F; Fig. 1b, lane 2). The differences in the extent of α -factor release into the MSS in the absence or presence of cytosol were comparable to the levels of transport activity observed with the basic two-stage assay conditions in Fig. 1a. When Sec7-Ab were introduced to a complete transport reaction, no outer-chain glycosylated α -factor was detected in the MSS. But the amount of α -factor released into the MSS was unaffected by the presence of Sec7-Ab, as seen by the increased density of CG- α F in lane 3 compared with lane 2 (Fig. 1b). These data suggested that transport vesicle budding from the ER was not blocked by the addition of Sec7-Ab, but subsequent steps in the transport process were disrupted.

Sec7p is recovered in roughly equivalent amounts from soluble and sedimentable fractions of wild-type yeast lysates²⁰. Immunofluorescence experiments in wild-type cells revealed that the highest concentration of Sec7p was found in association with the yeast Golgi apparatus²⁰. Therefore, we tested whether the Golgi membranes were rendered incompetent as a target

compartment by the binding of the antibodies. The addition of increasing concentrations of fresh Golgi membranes from an untreated fraction was unable to restore transport activity in the presence of Sec7-Ab. By contrast, the addition of cytosolic proteins comparable to the lowest concentration of membrane proteins reversed the inhibition of transport activity by Sec7-Ab (data not shown). This indicated that the Golgi membranes were not disabled by the antibodies.

Vesicles containing α -factor intermediates that had apparently budded from the ER seemed to lose competence for subsequent transport activity owing to the introduction of Sec7-Ab. The presence of Sec7-Ab (and therefore Sec7p) on the vesicle surface was tested by immunologically recovering the transport vesicles containing the radiolabelled α -factor cargo. The transport assay was done in the presence of Sec7-Ab, the IgG fraction of preimmune sera or no antibodies. Protein A-Sepharose was mixed with the MSS from each assay condition (containing virtually identical concentrations of α -factor, see Fig. 2a; MSS, lanes 1–3). The amount of α -factor recovered from the MSS with protein A-Sepharose in the assays treated with Sec7-Ab was about five-fold greater than that seen with either preimmune IgG or no antibodies added to the transport reaction (Fig. 2a, lanes 4–6, quantitation shown in Fig. 2b). Neither affinity-purified HMG-CoA reductase nor Kex2 protein antibodies (which recognize the cytoplasmic domain of transmembrane proteins of the yeast ER and Golgi apparatus, respectively^{20,23}) could substitute for Sec7-Ab in the accumulation or retrieval of these transport vesicles. Furthermore, western blot analysis did not detect transmembrane ER proteins (HMG-CoA

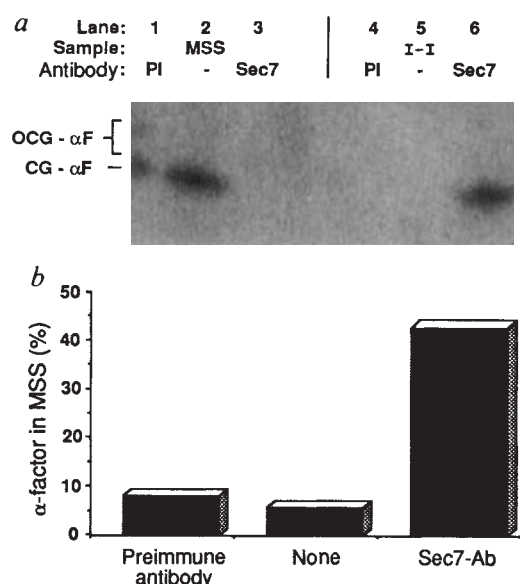
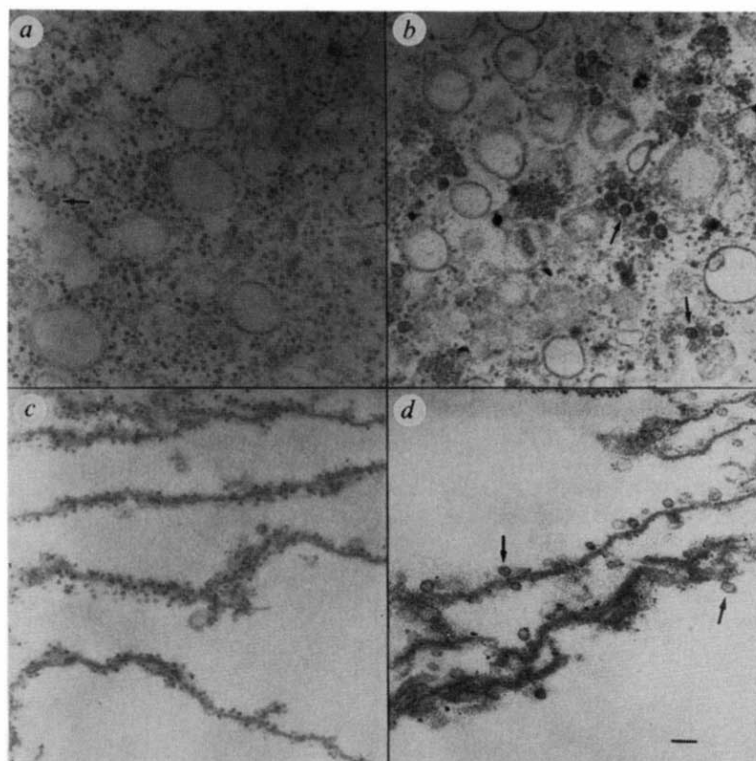


FIG. 2 Immuno-isolation of Sec7p-coated vesicles from the cell-free assay. **a**, The extent of the transport cargo α -factor immuno-isolated (I-I; lanes 4–6) with protein A-Sepharose from the MSS (lanes 1–3) of cell-free reactions done in the presence of IgG of preimmune sera (PI; lanes 1, 4), no antisera (–; lanes 2, 5), or Sec7-Ab (Sec7; lanes 3, 6). **b**, Quantitation of the immuno-isolated radiolabelled α -factor as a function of the total material in the MSS. Equivalent amounts of α -factor cargo were released into the MSS of all three reactions. Over 90% of the total α -factor in the MSS from these experiments was protease-resistant in the absence, but not in the presence of detergent (data not shown). This suggested that the α -factor was localized in the lumen of membrane compartments.

METHODS. The MSS was prepared as described in the legend to Fig. 1. The second stage of the transport reaction was done in the presence of wild-type cytosol plus preimmune IgG, no antisera or Sec7-Ab, as described in the legend to Fig. 1. The reaction was incubated for 15 min at 20 °C, then the mixture was sedimented as before. Aliquots of the MSS fraction (twice the volume loaded directly on the gels) were mixed with protein A-Sepharose in assay buffer for 2 h at 25 °C or overnight at 4 °C. The beads were allowed to settle, washed four times with buffer 88, then solubilized in SDS-sample buffer. The amount of radiolabelled α -factor that coprecipitated with the protein A-Sepharose beads was quantified by two different methods. The radioactivity in the gel was quantified using a Phosphor-imager (Molecular Dynamics). Conversely, concanavalin A-Sepharose binding, washing and scintillation counting of the SDS-solubilized immuno-isolated precipitates was also used. The values represent the fraction of α -factor polypeptides precipitated from the MSS, compared with the total transported α -factor protein from a complete reaction. Similar results were reproduced in four separate experiments.

FIG. 3 Morphology of Sec7p-coated transport vesicles in the cell-free assay. MSS from the cell-free assay treated with the IgG fraction from preimmune sera (a) and the immuno-isolated fraction from this MSS (c). The corresponding samples from cell-free assays done in the presence of Sec7-Ab (b, MSS; d, immuno-isolated fraction). The visible aggregation of the 40–50 nm diameter vesicles in the Sec7-Ab-treated MSS may be because of crosslinking by the antibody. Similar vesicles in the control MSS were more uniformly dispersed. The fibres observed in the immuno-isolated fractions represent the thin shell of nonmagnetic material which dissociated from the bead surface during EM preparation²⁷. The small black particles associated with the fibres are magnetic from the bead core. Scale bar, 100 nm.

METHODS. The MSS was prepared and used for immuno-isolation with magnetic beads (kindly provided by J. Ugelstad, University of Trondheim, Norway; and from Dynal Inc. (Great Neck, NY, USA)). The immuno-adsorbent was treated as described in the legend to Fig. 2, except that a magnetic block was used in retrieving the beads during the washes. The fractions were fixed in suspension with 1% glutaraldehyde in cacodylate buffer, stained with 2% tannic acid, postfixed in 2% osmium, dehydrated and embedded in Spurr's resin²⁵. Thin sections were stained with uranyl acetate and examined in a Philips CM-10 microscope.



reductase²³ and Sec61 proteins²⁴) in the immuno-isolated vesicle fraction, whereas these antigens were present in the MSS (data not shown). This provides evidence that the vesicles are not fragmented derivatives of the donor ER membranes.

Transport vesicles with a (non-clathrin) coat have been identified by electron microscopy from mammalian cell-free assays disrupted with the nucleotide analogue, GTP- γ -S (ref. 25). Vesicles accumulated in yeast cell-free assays have not been previously characterized by EM. Sec7p-containing vesicles were immuno-isolated from the MSS using magnetic beads with covalently coupled mouse anti-rabbit antibodies, and visualized by electron microscopy. Vesicles (40–50 nm diameter) were recovered on the immuno-adsorbent when the cell-free assays were incubated with Sec7-Ab, but not the preimmune IgG (Fig. 3c, d). A population of vesicles with similar morphology was present in the Sec7-Ab treated MSS. The tannic acid stain was enriched around the vesicle membranes compared with other membrane compartments in the same field. In the MSS from control samples, these vesicles were evident but not as abundant (Fig. 3a, b). Quantitation of the 40–50 nm vesicles from the MSS that were bound to the immuno-adsorbent revealed a 10–20-fold enrichment over the controls (Table 1). The larger (>100 nm) less densely stained vesicles on the immuno-adsorbent were far less abundant, and not significantly enriched by comparison with the control samples. Vesicles of ~50 nm diameter accumulate *in vivo* by the action of yeast *sec* mutations that block membrane traffic from the ER (refs 15–17). Because Sec7p is required for multiple stages of membrane traffic through the yeast secretory pathway¹⁸, the Sec7p-associated vesicles visualized by electron microscopy may include intermediates from stages other than transport from the ER. The present cell-free assay must be modified to distinguish these intermediates biochemically.

The cytosolic protein Sec7p is associated with transport vesicles in transit from the ER to the Golgi apparatus. In contrast to the phenotype of exaggerated organelle morphology in the mutant strain *in vivo*^{18,26}, the Sec7-Ab do not prevent the budding of vesicles *in vitro*, but disrupt subsequent events leading to the fusion of vesicle intermediates with the acceptor compartments. Therefore, the transport vesicles could be immuno-isolated for

biochemical and morphological characterization. Other laboratories have purified transport vesicles from yeast cell-free assays^{11,12,14}. Vesicles have been accumulated in the presence of Ypt1 antibodies^{13,14}, but we have no evidence whether the two vesicle types exhibiting Sec7p or Ypt1p on their surface are identical or represent different steps from budding to fusion. The dense staining of the vesicle membranes suggests that these vesicles may have a (non-clathrin) protein coat similar to that observed in mammalian assays. The molecular details for the association of Sec7p, and perhaps other cytosolic proteins into this coat structure, and its function in membrane traffic are currently under investigation. □

Received 22 July; accepted 11 October 1991.

- Warren, G. *Nature* **345**, 382–384 (1990).
- Rothman, J. E. & Orci, L. *FASEB J.* **4**, 1460–1468 (1990).
- Melanson, P., Franzusoff, A. & Howell, K. *Trends Cell Biol.* **114**, 219–229 (1991).
- Serafini, T. et al. *Nature* **349**, 215–220 (1991).
- Salamero, J., Szul, E. S. & Howell, K. E. *Proc. natn. Acad. Sci. U.S.A.* **87**, 7717–7721 (1990).
- Tooze, S. & Huttner, W. B. *Cell* **60**, 837–847 (1990).
- deCurtis, I. & Simons, K. *Cell* **58**, 719–727 (1989).
- Balch, W. E. *Trends biochem. Sci.* **15**, 473–477 (1990).
- Paulik, M., Nowack, D. D. & Morrè, D. J. *J. biol. Chem.* **263**, 17738–17748 (1988).
- Lodish, H. F., Kong, N., Hirani, S. & Rasmussen, J. J. *Cell Biol.* **104**, 221–230 (1987).
- Groesch, M. E., Ruohola, H., Bacon, R., Rossi, G. & Ferro-Novick, S. *J. Cell Biol.* **111**, 45–53 (1990).
- Rexach, M. & Schekman, R. *J. Cell Biol.* **114**, 219–229 (1991).
- Baker, D., Wuestehube, L., Schekman, R., Botstein, D. & Segev, N. *Proc. natn. Acad. Sci. U.S.A.* **87**, 355–359 (1990).
- Segev, N. *Science* **252**, 1553–1556 (1991).
- Kaiser, C. A. & Schekman, R. *Cell* **61**, 723–733 (1990).
- Newman, A. P. & Ferro-Novick, S. *J. Cell Biol.* **105**, 1587–1594 (1987).
- Novick, P., Field, C. & Schekman, R. *Cell* **21**, 205–215 (1980).
- Franzusoff, A. & Schekman, R. *EMBO J.* **8**, 2695–2702 (1989).
- Achstetter, T., Franzusoff, A., Field, C. & Schekman, R. *J. biol. Chem.* **263**, 11711–11717 (1988).
- Franzusoff, A., Redding, K. R., Crosby, J., Fuller, R. S. & Schekman, R. *J. Cell Biol.* **112**, 27–37 (1991).
- Baker, D., Hicke, L., Rexach, M., Schleyer, M. & Schekman, R. *Cell* **54**, 335–344 (1988).
- Ruohola, H., Kabaceni, A. K. & Ferro-Novick, S. *J. Cell Biol.* **107**, 1465–1476 (1988).
- Wright, R., Basson, M., D'Ari, L. & Rine, J. *J. Cell Biol.* **107**, 101–114 (1988).
- Deshaies, R. J., Sanders, S. L., Feldheim, D. A. & Schekman, R. *Nature* **349**, 806–808 (1991).
- Malhotra, V., Serafini, T., Orci, L., Shepherd, J. C. & Rothman, J. E. *Cell* **58**, 329–336 (1989).
- Novick, P., Ferro, S. & Schekman, R. *Cell* **25**, 461–469 (1981).
- Howell, K. E., Schmid, R., Ugelstad, J. & Gruenberg, J. *Meth. Cell Biol.* **31**, 265–292 (1989).

ACKNOWLEDGEMENTS. We thank M. L. Kruse, M. Ladinsky and V. Miller for their assistance in the preparation of the samples for microscopy. Antisera specific for Kex2, Sec61 and HMG-CoA reductase proteins were provided by the laboratories of R. Fuller, R. Schekman and R. Wright, respectively. Use of the Hepatobiliary Center Core facilities in Membrane Biochemistry and Biophysics is gratefully acknowledged. This work was supported by grants from the Council for Tobacco Research, USA (to A.F.) and the NIH (to K.E.H.).

Independent binding of the retinoblastoma protein and p107 to the transcription factor E2F

Liang Cao, Barbara Faha, Marlene Dembski, Li-Huei Tsai, Ed Harlow & Nicholas Dyson*

Massachusetts General Hospital Cancer Center, Building 149, 13th Street, Charlestown, Massachusetts 02129, USA

* To whom correspondence should be addressed

THE cellular protein p107 and the retinoblastoma protein (pRB) have many features in common. Most strikingly, they contain homologous protein domains that mediate interaction with the oncoproteins of several small DNA tumour viruses, including adenovirus E1A and SV40 large-T antigen¹⁻⁹. In cells that do not contain these viral oncoproteins, pRB interacts with the cellular transcription factor E2F (refs 10-12) or a related protein termed DRTF1 (ref. 13). E2F associates with a form of pRB that is found primarily in G1 cells¹⁰. It seems that the E2F-pRB complex dissociates near the G1-S boundary before the initiation of S phase, releasing free E2F and apparently stimulating the ability

of E2F to activate transcription¹⁴. Cells that express E1A have no or little pRB-E2F complex, presumably because of the association of E1A with pRB^{10,15}. During S phase, E2F forms a second complex that contains cyclin A but apparently lacks pRB¹⁵. Here, we report that p107 is found in the cyclin A/E2F complex and that this complex also contains p33^{cdk2}. These observations suggest that p107 and pRB cooperate in the regulation of E2F activity, each affecting different stages of the cell cycle. Thus, by binding to pRB and p107, E1A and large-T antigen target two distinct aspects of E2F regulation.

We were prompted to examine the possibility that p107 may be involved in E2F regulation by four pieces of information: the observation that pRB associates with E2F¹⁰⁻¹²; the extensive similarity between pRB and p107¹⁻⁹; the description of independent E2F/pRB and E2F/cyclin A¹⁵ complexes; and that p107 and cyclin A form stable protein complexes in the absence of E1A^{16,17}. We raised polyclonal antibodies that were specific for p107 and that immunoprecipitated p107 polypeptides synthesized by *in vitro* translation of complementary RNA prepared from the p107 cDNA⁹ (Fig. 1a). The anti-p107 sera specifically recognized a protein of relative molecular mass (*M_r*, 107K) from labelled human cells but showed no crossreactivity to pRB (Fig. 1b). The 107K protein recognized by the anti-p107 sera comigrated with E1A-associated p107 protein isolated from human 293 cells (data not shown). Comparison of the 107K

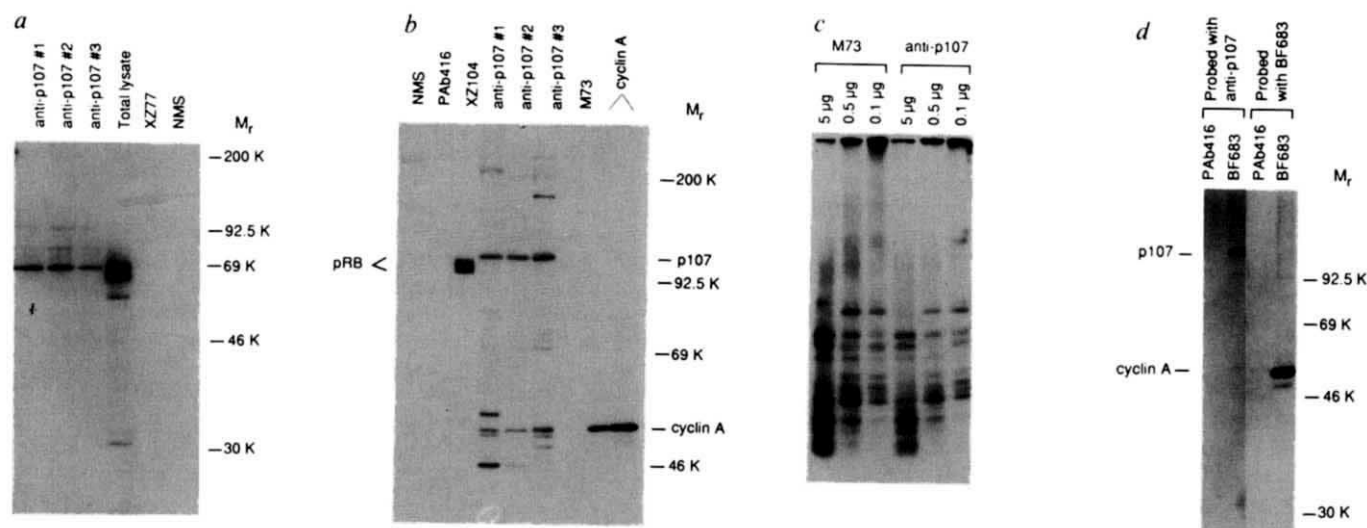


FIG. 1 Characterization of anti-p107 antibodies. *a*, Immunoprecipitation of polypeptides synthesized from the p107 cDNA clone. Sera were collected from three mice immunized with a glutathione-S-transferase (GST)-p107 fusion protein and used to immunoprecipitate radiolabelled p107 polypeptides. Normal mouse serum (NMS) and an anti-pRB monoclonal antibody, XZ77 (ref. 24), were used as negative controls. The position of size markers is indicated. *b*, Anti-p107 sera immunoprecipitate a 107K protein from ML-1 cell lysates. Proteins were immunoprecipitated with the antibodies indicated. XZ104 is a pRB-specific monoclonal antibody²⁴ that immunoprecipitates the multiple phosphorylated forms of pRB. M73²⁵, PAb416²⁶ and normal mouse serum (NMS) are negative control antibodies. In addition to the 107K protein, the anti-p107 antibodies immunoprecipitate a protein that comigrates with cyclin A. The position of size markers is indicated. *c*, Partial proteolytic mapping of p107. 107K proteins isolated through immunoprecipitation of E1A or with the anti-p107 sera were digested with the varying amounts of V8 protease indicated. The partial proteolytic products produced from these two proteins are identical, indicating that these proteins are the same. *d*, Immunoblots probed with the anti-p107 antibodies. Immune complexes prepared from ML-1 cells with either a cyclin A antibody, BF683¹⁶, or a negative control antibody, PAb416²⁶, were western blotted. Identical filters were probed with anti-p107 serum or with the cyclin A monoclonal antibody (BF683). Although cyclin A was detected by BF683, the anti-p107 serum was specific for the smaller amount of p107 that is present in cyclin A immunoprecipitates.

METHODS. *a*, ³⁵S-methionine-labelled polypeptides were produced by *in vitro*

transcription of the human p107 cDNA clone⁹ and translation of the cRNA in rabbit reticulocyte lysates (Promega)²⁷. Each serum (1 µl) and 50 µl of XZ77 were used for immunoprecipitation. Immune complexes were separated on a 10% SDS-polyacrylamide gel²⁸ and detected by fluorography²⁹. *b*, Cyclin A was produced by *in vitro* transcription and translation of the human cDNA clone. ML-1 cells were metabolically labelled with ³⁵S-methionine and lysed as described previously²⁴. Immunoprecipitation was done with 4 µl anti-p107 polyclonal antibodies or with 50 µl monoclonal antibody. Immune complexes were separated on an 8% SDS-polyacrylamide gel²⁸ and labelled proteins detected by fluorography²⁹. *c*, 293 cells were labelled with ³⁵S-methionine, and proteins immunoprecipitated with M73, a monoclonal antibody against E1A²⁵. ML-1 cells were similarly prepared and proteins immunoprecipitated with anti-p107 serum number 1. Immune complexes were separated on a 6% SDS-polyacrylamide gel²⁸. The gel was dried, and proteins were located by autoradiography. V8 partial proteolysis was done using the procedure described by Cleveland *et al.*³⁰. Slices containing the 107K proteins were excised from the gel and digested with *S. aureus* V8 protease. Digested proteins were separated on a 15% SDS-polyacrylamide gel²⁸ and visualized by fluorography²⁹. *d*, Immune complexes were collected on antibodies covalently coupled to protein A-Sepharose beads. Proteins separated on an 8% SDS-polyacrylamide gel were transferred to nitrocellulose³¹. The positions of specifically bound antibodies were detected with a rabbit anti-mouse alkaline phosphatase conjugate (BRL) and visualized as previously described³².

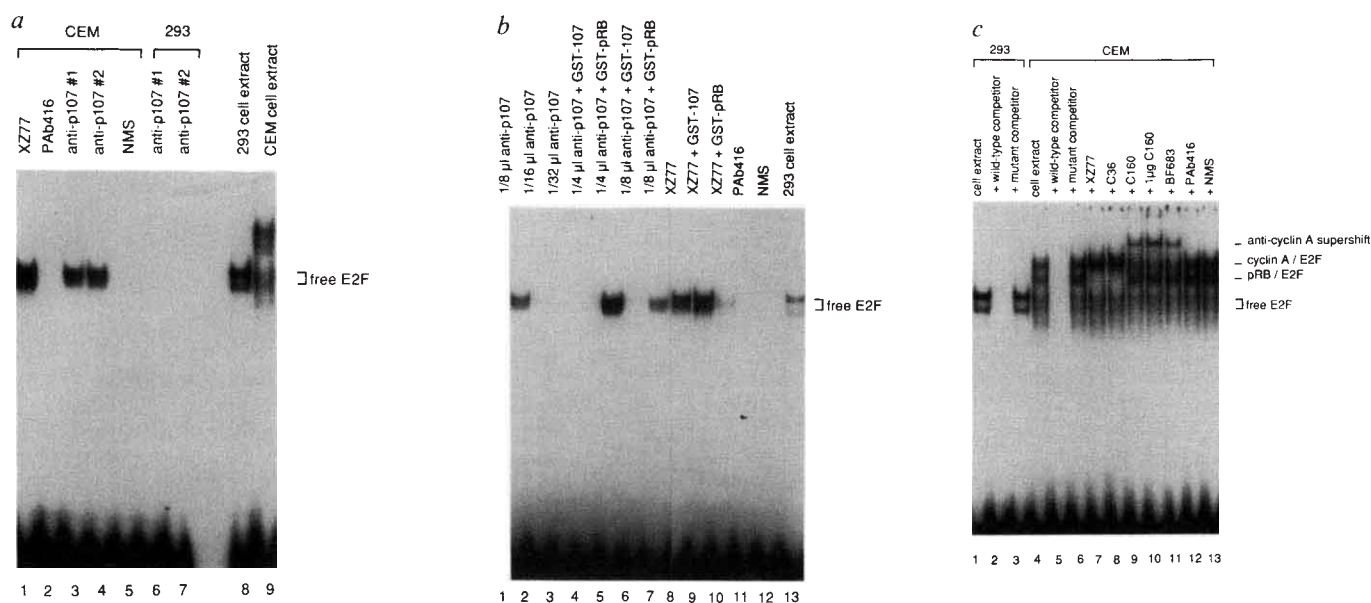


FIG. 2 Association of p107 with E2F. **a**, Anti-p107 antibodies specifically coimmunoprecipitate E2F. Cell extracts from CEM cells and 293 cells were immunoprecipitated with the antibodies indicated and immune complexes were assayed E2F activity. Anti-p107 sera from two mice (numbers 1 and 2) were tested on CEM cell extracts (lanes 3 and 4) and 293 cell extracts (lanes 6 and 7). E1A-containing 293 cells have almost exclusively the free forms of E2F and show no E2F activity associated with p107 (lanes 6, 7). The pRB monoclonal antibody (XZ77)²⁴ provides a positive control (lane 1) for the E2F-pRB complex in CEM cells. PAb416²⁶ and normal mouse serum (NMS, lane 5) are negative controls. Bandshifts obtained with total CEM and 293 cell extracts are shown (lanes 8 and 9). **b**, The anti-p107 antibody is blocked with purified GST-p107. Before immunoprecipitation, antibodies were incubated with purified GST fusion proteins. Lanes 1–3 show a titration of anti-p107 sera in the absence of GST-fusion protein. E2F coprecipitation by the anti-p107 serum was inhibited by preincubation with GST-p107 (lanes 4 and 6) but not with GST-pRB³³ (lanes 5 and 7). Conversely, E2F coprecipitation by anti-pRB monoclonal antibody was inhibited only by GST-pRB (lane 10). PAb416²⁶ and normal mouse serum (NMS) are negative controls (lanes 11 and 12). **c**, Identification of pRB- and cyclin A-containing E2F complexes in CEM cell extracts. Lanes 1–6 show E2F activity in 293 and CEM cell extracts. Specific E2F complexes were competed away by an excess of unlabelled competitor DNA containing the wild-type sequence (lanes 2 and 5) but not by a mutated competitor (lanes 3 and 6). Distinct forms of E2F are identified by the addition of monoclonal antibodies to the binding reactions (lanes 7–11). Two anti-pRB monoclonal antibodies, XZ77²⁴ and C36¹ (lanes 7 and 8) specifically abolished or supershifted a single band of E2F (labelled E2F-pRB). Two anti-cyclin A monoclonal antibodies, C160¹⁹ and BF683¹⁶ (lanes 9–11) supershifted most of the higher band (labelled E2F/cyclin A) to produce a new band (labelled anti-cyclin A supershift). Pab416 and normal mouse serum (NMS) were negative controls (lanes 12, 13). A complete shift of the slowest migrating E2F complex by monoclonal antibodies to cyclin A was not seen even when affinity purified antibody was added in great excess (lane 10). The reasons for this are unclear. **d**, An E2F complex contains both p107 and cyclin A. Lanes 1–4, controls that identify the respective forms of E2F as described in Fig. 2c. Lanes 5–14, bandshift reactions using CEM cell extracts. The effects of anti-p107 antibodies on E2F complexes was examined in the absence (lanes 5–9) or the presence (lanes 10–14) of the cyclin A monoclonal antibody (C160) which generates a supershifted E2F/cyclin A complex (lane 10). Addition of anti-p107 serum alone selectively eliminated the slowest form of E2F (lanes 6–8). In the presence of the anti-cyclin A antibody (C160), the anti-p107

serum also selectively eliminated the supershifted E2F/cyclin A complex. Lanes 5 and 10 contain no anti-p107 antibody. Lanes 9 and 14 show negative controls containing normal mouse serum (NMS). **e**, E2F activity associated with p107 in extracts prepared from multiple cell lines. E2F activity was assayed in immune complexes prepared with anti-p107 serum (lanes 1) normal mouse serum (lanes 2), or the anti-pRB antibody (XZ77) (lanes 3) from a panel of human cell lines. 293 is a human kidney cell line transformed with E1A and E1B, CEM and ML-1 are derived from human leukaemias, Saos-2, KHOS-240S and U20S are osteosarcoma cell lines. Saos-2 cells produce a truncated pRB that is located in the cytoplasm¹⁸. An additional lane containing 5 μ g total 293 cell extract indicates the position of free E2F. **METHODS.** **a**, Cell extract (200 μ g) was used per immunoprecipitation and immune complexes were incubated with 0.8% deoxycholate to release associated E2F activity. NP40 was added to a final concentration of 1.5% and samples were subjected to gel shift assay with a ³²P-labelled E2F oligonucleotide probe^{15,17}. About 5–10 μ g of proteins were used directly for the gel shift assay in a binding buffer containing 20 mM HEPES, pH 7.4, 40 mM KCl, 1 mM MgCl₂, 0.1 mM EDTA, 0.1% NP40. The E2F oligonucleotide probe was ATTTAAGTTTCGATCCCTTTCTCAA. **b**, Preincubation was done with 20 μ g purified GST fusion protein for 2 h before immunoprecipitation. Immunoprecipitation and gel shift analysis were done as described above. **c**, Antibodies were added as 1 μ l of tissue culture supernatant, polyclonal serum, or antibody purified from ascites (C160, lane 10). Gel shift analysis was done as described above. Competitor oligonucleotides were added in 1,000-fold excess; the mutant oligonucleotide used was ATTTAAGTTTCGATCCCTTTCTCAA. **e**, Cell extract (100 μ g) was used for immunoprecipitation. Immune complexes were treated with deoxycholate and assayed for releasable E2F gel shift activity as described in Fig. 2a.

protein with E1A-associated p107 by partial proteolytic mapping with *Staphylococcus aureus* V8 protease confirmed that it was authentic p107 (Fig. 1c).

A few proteins were precipitated by the anti-p107 antibodies in addition to p107. The most abundant of these comigrated with human cyclin A. This is not surprising because p107/cyclin A complexes have been identified previously using immunological reagents to cyclin A¹⁶. As predicted by these earlier observations, anti-p107 antibodies did not seem to recognize cyclin A directly. When used to probe western blots, the anti-p107 antibodies detected p107, but failed to detect cyclin A (Fig. 1d). Similar results were obtained using radiolabelled cyclin A and p107 synthesized by *in vitro* translation (data not shown).

Next, we used the anti-p107 antibodies to determine whether p107 is associated with E2F. Cell extracts were prepared from human cells that either contain (293) or lack E1A (CEM) and were immunoprecipitated with anti-p107 sera or with control antibodies. The immune complexes were assayed for E2F activity using a standard bandshift assay (Fig. 2a). The anti-p107 antibodies specifically coprecipitated E2F activity from CEM extracts, but no activity was associated when E1A-containing 293 cell extracts were used. Furthermore, E2F coprecipitation by the anti-p107 antibodies was abolished by preincubating the antisera with purified GST-p107 protein but not by preincubation with GST-pRB (Fig. 2b).

A bandshift assay was used to determine which of the various E2F complexes contained p107. Extracts from CEM cells gave four specific major bands as reported previously^{10,15}. The two fastest migrating forms represent free E2F; the lower band is thought to be a partial breakdown product of E2F. E2F/cyclin A complexes and E2F-pRB complexes were identified using monoclonal antibodies specific for pRB and cyclin A that either

selectively eliminated or supershifted specific DNA-bound complexes (Fig. 2c). Addition of anti-p107 sera to the E2F bandshift selectively eliminated the slow-migrating E2F complex. This was the same band of E2F that was supershifted by antibodies to cyclin A (Fig. 2d). Because the cyclin A-monoclonal antibodies induced a supershift of the cyclin A/E2F complex, we added anti-p107 serum and anti-cyclin A antibodies together to the binding reaction to investigate whether both reagents could recognize the same E2F complex. The supershifted form of cyclin A/E2F that contained the anti-cyclin A antibody was selectively eliminated by anti-p107 sera (Fig. 2d). This observation strongly argues that cyclin A and p107 are present in the same E2F complex.

If p107 has an important role in the regulation of E2F activity, then its association with E2F should be ubiquitous. Cell extracts were made from six different cell lines and anti-pRB or anti-p107 immunoprecipitates were assayed for associated E2F. In four lines (CEM, ML1, KHOS-240S, U20S) E2F activity was seen in immunoprecipitates prepared with both anti-RB and anti-p107 antibodies, but not with control antibodies (Fig. 2e). When Saos-2 cells (containing mutated RB-1¹⁸) were used, E2F was seen in anti-p107 immunoprecipitates but not in anti-pRB immunoprecipitates. This result further suggests that E2F association with p107 is independent of its association with pRB. As shown earlier, no E2F was associated with p107 or pRB in the E1A-containing cell (293) extracts.

Human cyclin A associates with two distinct kinase subunits, p34^{cdc2} (B.F., unpublished results) and p33^{cdc2} (refs 19–22 and B.F., unpublished results), both of which are thought to be key regulators of cell-cycle progression. We therefore also used antibodies to these proteins to determine whether either of these kinases is associated with E2F. Anti-cdk2 serum²² coprecipitated

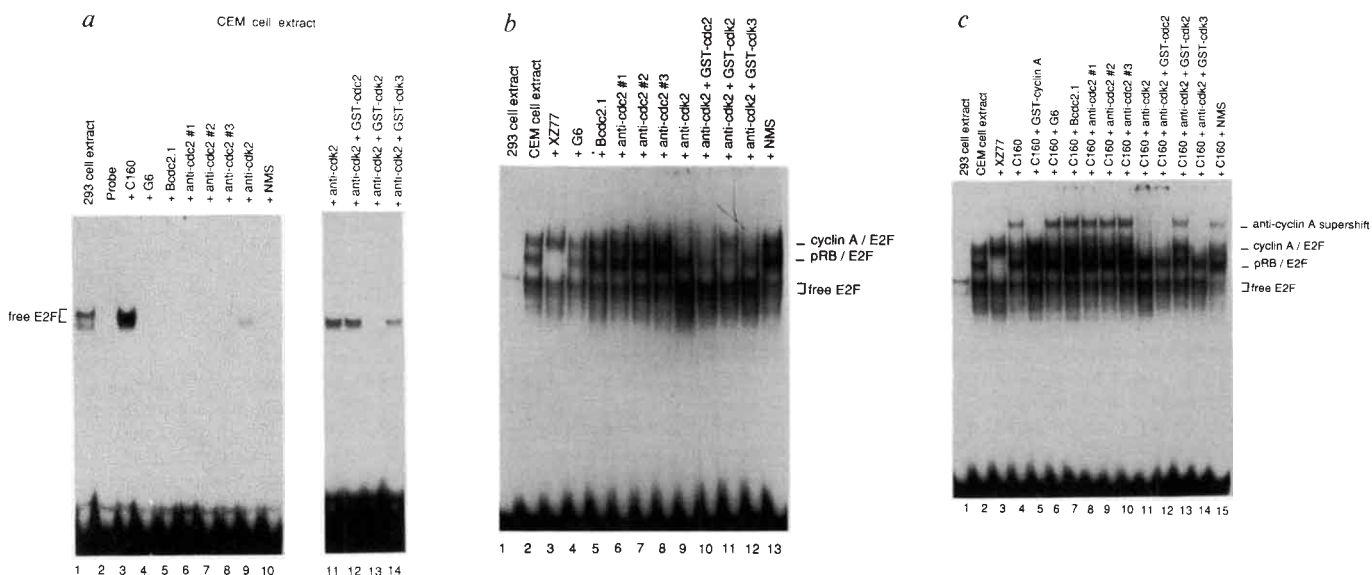
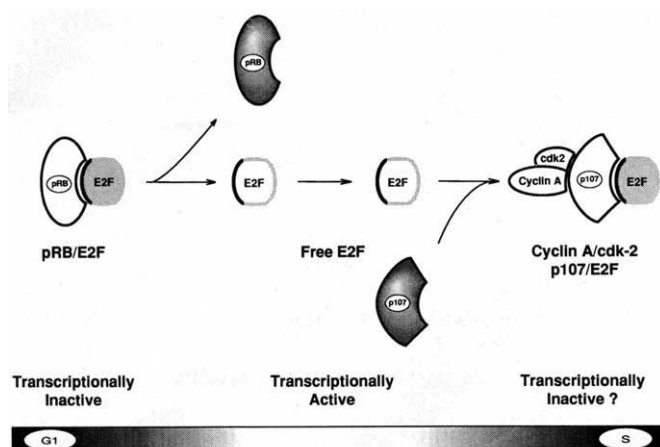


FIG. 3 An E2F complex containing cyclin A and p33^{cdc2}. **a**, A p33^{cdc2} antisera specifically coprecipitates E2F. CEM cell extracts were immunoprecipitated with the antibodies indicated and immune complexes were assayed for E2F activity as described in Fig. 2a. 293 cell extract provides a marker for free E2F. C160 is a monoclonal antibody to cyclin A¹⁹. G6 and Bcd2.1 are antibodies to human p34^{cdc2} (refs 34, 22). Mouse polyclonal antibodies to p34^{cdc2} (lanes 6–8) and p33^{cdc2} (lane 9) were generated as previously described²². Normal mouse serum (NMS) was added as a negative control. Coimmunoprecipitation of E2F by the anti-cdk2 serum was inhibited by preincubating the serum with 1 μ g purified GST-cdk2 (lane 13) but not by incubation with GST-cdc2 or GST-cdk3 (lanes 12 and 14). In the family of cdc2-related kinases, cdc2 and cdk3 are the proteins that are most similar to cdk2³⁵. **b**, Anti-cdk2 serum specifically eliminates the slow-migrating E2F complex. The same panel of antibodies to p34^{cdc2} and p33^{cdc2} were assayed as described in Fig. 2a. Lanes 1 and 2, E2F complexes that are present in untreated 293 and CEM cell extracts. XZ77, a monoclonal

antibody to pRB, was added in lane 3 to identify the E2F-pRB complex. Anti-cdk2 serum was used directly (lane 9) or after preincubation with 1 μ g purified GST-cdc2, GST-cdk2 and GST-cdk3 (lanes 10, 11 and 12). Anti-cdk2 specifically eliminated the slowest migrating form of E2F (lane 9) but this elimination was blocked by incubation with GST-cdk2 protein (lane 11). **c**, Cyclin A and p33^{cdc2} are in the same E2F complex. The panel of antibodies to p33^{cdc2} and p34^{cdc2} (see above) were added to an E2F gel shift together with the anti-cyclin A antibody, C160¹⁹. C160 induced a supershift of the slowest migrating form of E2F (lane 4), but preincubation of the antibody with GST/cyclin A abolished this effect (lane 5). The C160 induced supershift was selectively eliminated by addition of anti-cdk2 serum (lane 11) but not by the addition of any of the antibodies to p34^{cdc2} (lanes 6–10). Preincubation of the anti-cdk2 serum with GST-cdk2 abolished the activity of the serum (lane 13), but preincubation with GST-cdc2 or GST-cdk3 had no effect (lanes 12 and 14).



E2F activity but a panel of polyclonal and monoclonal antibodies to cdk2 showed no associated activity (Fig. 3a). E2F coprecipitation by the anti-cdk2 serum was abolished by preincubating the antisera with purified GST-cdk2 fusion protein but not by preincubation with homologous proteins (GST-cdc2 or GST-cdk3) indicating that E2F association is specific to cdk2. When added to E2F bandshift experiments, the anti-cdk2 antibodies gave results identical to those obtained with anti-p107 sera. Anti-cdk2 serum specifically eliminated the slowest migrating form of E2F (Fig. 3b). When anti-cdk2 serum was added in addition to an anti-cyclin A monoclonal antibody, the anti-cdk2 antibodies specifically eliminated the supershifted form of E2F that contained E2F, cyclin A and the cyclin A antibody (Fig. 3c). This effect was specific to cdk2 because the addition of anti-cdc2 antibodies had no effect on the E2F bandshift and the activity of the anti-cdk2 serum was specifically blocked by preincubation with purified GST-cdk2. Taken together with the data described above these results define an E2F complex that contains stoichiometric amounts of E2F, cyclin A, p107 and cdk2.

The previous observations that E2F is independently associated with pRB and cyclin A were perplexing, as these cellular proteins were not known to contain similar structures. The demonstration that the RB-related protein p107 is a major component of the E2F/cyclin A complex clarifies this situation. This finding suggests that the similarity between p107 and pRB provides the structural basis for interaction with E2F and indicates that cyclin A is associated with E2F indirectly through its interaction with p107. The initial identification of an E2F/cyclin A complex was suggestive of a novel function for a cyclin. But that p33^{cdk2} is present in the E2F/cyclin A/p107 complex strongly indicates that cyclin A serves its conventional role as a regulator of kinase activity. As a multimeric kinase with specific DNA binding activity, the identification of the E2F/p107/cyclin A/p33^{cdk2} complex raises many further questions. The most puzzling question is what is the role of a cell cycle regulating kinase that can be specifically localized to a particular DNA sequence?

It is appealing to speculate that pRB and p107 serve similar functions in their association with E2F. Free E2F has been correlated with transcriptional activation of E2F-driven promoters, and it has been suggested that both the E2F-pRB and E2F/cyclin A complexes may be transcriptionally inactive^{14,15}. Studies of human and mouse cells suggest that E2F associates with pRB and cyclin A at different times of the cell cycle^{10,15}. An attractive model (Fig. 4) is that pRB and p107 act as confiners for E2F activity, signalling the timing of the activation and inactivation of E2F function and perhaps coordinating the expression of genes required for DNA synthesis²³. Thus by

binding to pRB and p107, E1A may use a single motif to target two points of E2F regulation. □

Received 30 October; accepted 16 December 1991.

- Whyte, P. *et al.* *Nature* **334**, 124-129 (1988).
- DeCaprio, J. A. *et al.* *Cell* **54**, 275-283 (1988).
- Whyte, P., Williamson, N. M. & Harlow, E. *Cell* **56**, 67-75 (1989).
- Dyson, N., Buchkovich, K., Whyte, P. & Harlow, E. *Cell* **58**, 249-255 (1989).
- Ewen, M. E. *et al.* *Cell* **58**, 257-267 (1989).
- Hu, Q., Dyson, N. & Harlow, E. *EMBO J.* **9**, 1147-1155 (1990).
- Kaelin, W. G., Ewen, M. E. & Livingston, D. M. *Molec. cell Biol.* **10**, 3761-3769 (1990).
- Huang, S. *et al.* *EMBO J.* **9**, 1815-1822 (1990).
- Ewen, M. E., Xing, Y., Lawrence, J. B. & Livingston, D. M. *Cell* **66**, 1155-1164 (1991).
- Chellappan, S., Hiebert, S., Mudryj, M., Horowitz, J. M. & Nevins, J. R. *Cell* **65**, 1053-1061 (1991).
- Bagchi, S., Weinmann, R. & Raychaudhuri, P. *Cell* **65**, 1063-1072 (1991).
- Chittenden, T., Livingston, D. M. & Kaelin, W. G. *Cell* **65**, 1073-1082 (1991).
- Bandara, L. & La Thangue, N. *Nature* **351**, 494-497 (1991).
- Hiebert, S. W., Chellappan, S. P., Horowitz, J. M. & Nevins, J. R. *Genes Dev.* (in the press).
- Mudryj, M. *et al.* *Cell* **65**, 1243-1251 (1991).
- Faha, B., Ewen, M. E., Tsai, L.-H., Livingston, D. M. & Harlow, E. *Science* (in the press).
- Ewen, M. E., Faha, B., Harlow, E. & Livingston, D. M. *Science* (in the press).
- Shew, J.-Y. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **87**, 6-10 (1990).
- Giordano, A. *et al.* *Cell* **58**, 981-990 (1989).
- Pines, J. & Hunter, T. *Nature* **346**, 760-763 (1990).
- Giordano, A. *et al.* *Science* **253**, 1271-1275 (1991).
- Tsai, L.-H., Harlow, E. & Meyerson, M. *Nature* **353**, 174-177 (1991).
- Mudryj, M., Hiebert, S. W. & Nevins, J. R. *EMBO J.* **7**, 2179-2184 (1990).
- Hu, Q. *et al.* *Molec. cell Biol.* **11**, 5792-5799 (1991).
- Harlow, E., Franza, B. R. Jr & Schley, C. J. *Virology* **55**, 533-546 (1985).
- Harlow, E., Crawford, L., Pim, D. & Williamson, N. J. *Virology* **39**, 861-869 (1981).
- Krieg, P. A. & Melton, D. A. *Nucleic Acids Res.* **12**, 7057-7070 (1984).
- Laemmli, U. K. *Nature* **227**, 681-685 (1970).
- Bonner, W. M. & Laskey, R. A. *Eur. J. Biochem.* **46**, 83-88 (1974).
- Cleveland, D. W., Fischer, S. G., Kirschner, M. W. & Laemmli, U. K. *J. Biol. Chem.* **252**, 1102-1106 (1977).
- Towbin, H., Staehelin, T. & Gordon, J. *Proc. natn. Acad. Sci. U.S.A.* **76**, 4350-4354 (1979).
- Harlow, E. & Lane, D. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York, 1988).
- Kaelin, W. G. *et al.* *Cell* **64**, 521-532 (1991).
- Draetta, G. & Beach, D. *Cell* **54**, 17-26 (1988).
- Meyerson, M., Faha, B., Harlow, E. & Tsai, L.-H. *Cold Spring Harb. Symp. quant. Biol.* **58**, (in the press).

ACKNOWLEDGEMENTS. We thank J. Nevins for support and stimulating discussions, M. Ewen for the generous gifts of p107 constructs before publication of the p107 sequence, A. Yee for advice and assistance in setting up the gel shift system, A. Fattaey for assistance in the preparation of the antibody, M. Meyerson for the gift of purified GST-cdk3 protein, J. Bartek for the gift of Bcdc2.1 antibody, L. Zhu for some of the initial characterization of anti-p107 antibodies, E. Lees for the *in vitro* translated cyclin A, K. Helin, A. Fattaey, E. Lees and C.-L. Wu for reading the manuscript and our colleagues at the MGH Cancer Center for discussions. This work was supported by an institutional award to E.H. from the Massachusetts General Hospital Cancer Center. L.C. is a recipient of a Damon Runyon-Walter Winchell Cancer Research Fund Fellowship.

New yeast actin-like gene required late in the cell cycle

Etienne Schwob* & Robert P. Martin†

Laboratoire de Biochimie, Institut de Biologie Moléculaire et Cellulaire, Centre National de la Recherche Scientifique and Université Louis Pasteur, 15 rue René Descartes, 67084 Strasbourg, France

ACTIN, a major cytoskeletal component of all eukaryotic cells, is one of the most highly conserved proteins. It is involved in various cellular processes such as motility, cytoplasmic streaming, chromosome segregation and cytokinesis^{1,2}. The actin from the yeast *Saccharomyces cerevisiae*, encoded by the essential *ACT1* gene³⁻⁵, is 89% identical to mouse cytoplasmic actin and is involved in the organization and polarized growth of the cell surface⁶⁻⁸. We report here the characterization of *ACT2*, a previously undescribed yeast split gene encoding a putative protein (391 amino acids, relative molecular mass (M_r) 44,073) that is 47% identical to yeast actin. The requirement of the *ACT2* gene for vegetative growth of yeast cells and the existence of related genes in other eukaryotes indicate an important and conserved role for these actin-like proteins. Superimposition of the Act2 polypeptide onto the three-dimensional structure^{9,10} of known actins reveals that most of the divergence occurred in loops involved in actin polymerization,

* Present address: Research Institute of Molecular Pathology, A-1030 Vienna, Austria.

† To whom correspondence should be addressed.

DNase I and myosin binding, leaving the core domain mainly unaffected. To our knowledge, the Act2 protein from *S. cerevisiae* is the first highly divergent actin molecule described. Structural and physiological data suggest that the Act2 protein might have an important role in cytoskeletal reorganization during the cell cycle.

The *ACT2* gene was cloned from a yeast genomic expression library¹¹ with the neighbouring *PRP9* gene¹². An open reading frame starting 210 base pairs (bp) downstream of the termination codon of the *PRP9* gene was identified by its similarity to the highly conserved family of actins. Because yeast actin is encoded by a single gene (*ACT1*)^{3,4}, we named this gene *ACT2*. Figure 1 describes its physical map, complete nucleotide sequence and corresponding deduced translation product. Hybridization experiments indicate that the *ACT2* gene is unique in the yeast genome and is located on chromosome IV (data not shown). The yeast actin gene *ACT1* is located on chromosome VI (ref. 13) and is interrupted in its fourth codon by a 309-bp-long intervening sequence. Sequencing of the *ACT2* gene revealed an open reading frame of slightly higher coding capacity (391 versus 375 amino acids) which is interrupted by a 123-bp intron in the eighth codon (Fig. 1b). Northern analysis uncovered a polyadenylated *ACT2* transcript of about 1,250 nucleotides with a steady-state level similar to the *URA3* messenger RNA, which is moderately expressed (data not shown). Whereas the *ACT1* gene is highly expressed and yields an abundant protein (roughly 1% of yeast total cellular proteins¹⁴), we expect the *ACT2* gene product to be much less abundant.

In vivo, actin polymerizes into filaments (7-nm diameter) that make up, together with the microtubules and intermediate filaments, the cytoskeleton. In budding yeast, actin microfilaments can be seen as cables and cortical patches whose spatial organization varies during the cell cycle and reflects the asymmetry of growth^{6,7}. Yeast actin is required for vegetative growth⁵ and is involved in the maintenance of cell shape and the polar-

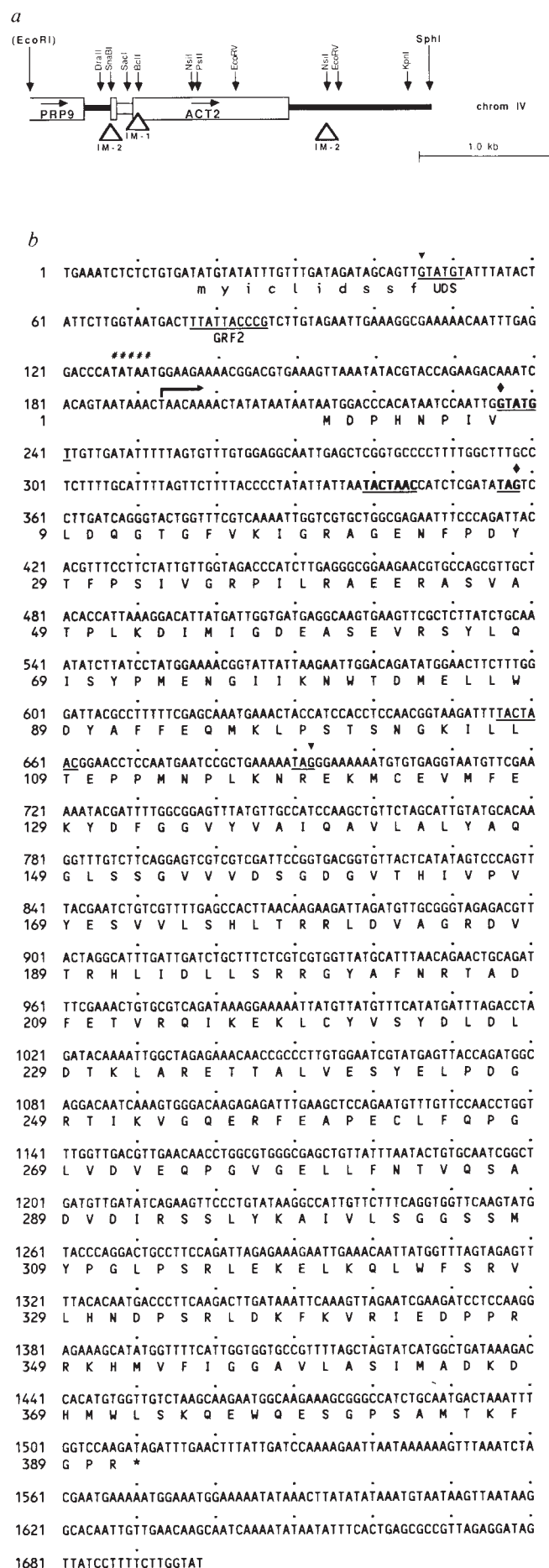


FIG. 1. The *ACT2* gene and its associated polypeptide. **a**, Physical map. The *EcoRI*-*SphI* genomic fragment contains the entire *ACT2* gene downstream of the *PRP9* gene¹². →, Orientation of the reading frames. Exons and intron are represented by open boxes and bar, respectively. Triangles, Sites used for *ACT2* disruption by insertion (IM-1) and replacement (IM-2) with the *URA3* marker gene. The *ACT2* gene is unique and located on chromosome IV as determined by Southern hybridization with restriction endonuclease digested DNA and OFAGE-blotted chromosomes (data not shown). The *EcoRI* site in parentheses is from the λ gt11 vector. **b**, Nucleotide and deduced amino-acid sequence of the *ACT2* gene. Numbering of nucleotides starts at the termination codon TGA of the upstream *PRP9* gene. The deduced amino-acid sequence is in standard single-letter code, with putative exon 1a in lower case. Major (♦) or alternative (▼) splice sites and their consensus sequences are in bold and underlined; UDS, upstream donor site. The existence of additional consensus sequences²⁵ for 5'-donor and branch-point splice sites in the *ACT2* gene led us to investigate the possibility of alternative splicing. The *Drall*-*SphI* subfragment in a centromeric plasmid, although lacking the putative exon 1a sequence, was able to complement an *act2::URA3* null mutation and therefore yielded an apparently functional *ACT2* gene (data not shown). The transcription start site (arrow) was determined as the nucleotide in position -21 relative to the initiation codon (data not shown), 116 and 66 bp downstream from a putative GRF2 binding-site²⁶ (GRF2) and TATA-box (# # # #), respectively.

METHODS. The λ gt11 clone containing the *ACT2* gene was isolated by a standard screening procedure¹¹ with polyclonal antibodies recognizing the *PRP9* gene product. Its insert was recombined in the M13mp18/19 vectors and nested deletions²⁷ were sequenced by the dideoxy termination method. DNA and protein sequences were analysed by the programs of the UWGCG package²⁸. The *ACT2* gene was disrupted with the *URA3* marker gene by both insertional and deletional mutagenesis using the one-step transplacement method²⁹ and standard yeast genetic techniques³⁰. Integration of the disrupted allele at the homologous *ACT2* locus was verified by Southern analysis of genomic DNA from diploid Ura⁺ transformants (data not shown). The transcription start site was determined by primer-extension³¹ using 5 μ g poly-(A)⁺ or 50 μ g total RNA and 0.5 pmol of a labelled oligonucleotide complementary to nucleotides 465-485.

ized growth of the cell surface during bud emergence⁸. We disrupted the *ACT2* gene with the *URA3* marker by insertional and replacement mutagenesis, followed by homologous recombination in a *ura3* diploid strain (Fig. 1a). The lack of any viable *Ura*⁺ spores after tetrad analysis demonstrated that the *ACT2* gene is essential (data not shown). Microscopic examination of the *act2::URA3* disruptants revealed in each case a single cell with a large bud, indicating that growth had arrested before the end of the first cell division (Fig. 2a). This is clearly different from the *ACT1* gene disruption where cells were unbudded or with a tiny bud⁵ and indicates that the Act2 protein is required for at least one vital step late in the cell cycle. Heterozygous *ACT2/act2::URA3* diploids show a marginal phenotype, with 5–20% (depending on the growth rate) of the cells having multiple (two to four) large buds, each of them containing a nucleus (Fig. 2b). This phenotype of delayed cytokinesis could be explained by a gene dosage effect, one copy of the functional *ACT2* gene not providing enough product to fulfill its biological role in fast-growing diploid cells. Additionally, the overexpression of the *ACT2* gene from a multicopy plasmid resulted in very slow growth and the accumulation of large rounded terminal cells typical of a disturbed cytoskeleton (data not shown). These phenotypes suggest that the expression of the Act2 protein is strictly controlled, as is the *ACT1* gene product¹⁵.

By low-stringency Southern hybridization with genomic DNA of several lower eukaryotes, we were able to detect *ACT2*-related sequences in *Schizosaccharomyces pombe*, *Aspergillus nidulans* and *Physarum polycephalum* (data not shown). This result, taken together with the recent identification of complementary DNAs coding for actin-like proteins in *Caenorhabditis elegans* (R. Waterston, St. Louis, personal communication), dog (D. Meyer, Los Angeles, personal communication), and human¹⁶, suggest that one or more classes of actin-like proteins are universally present in eukaryotes.

Figure 3 shows an alignment of the yeast Act2 and nematode actin-like protein with yeast and rabbit actin¹⁷. All actins identified (more than 40 isoforms from a variety of genetic phyla) are the same size (375 amino acids) and usually share >80% identical residues¹. For example, the yeast actin is 88% identical to the rabbit α -actin and up to 96% similar if conservative changes are also taken into account. Not a single pair of actin molecules has <70% identical (85% similar) residues. By contrast, the Act2 protein (391 amino acids) is larger than both yeast and rabbit α -actin and displays 'only' 47% identical (74% similar) residues. We therefore believe that the Act2 protein identifies a new class of actin-like proteins. The size difference between the Act2 protein and known actins results mainly from a 3 and an 11 amino-acid insertion at residues 42 and 320, respectively. All the amino-acid substitutions are scattered along the polypeptide chain. The nematode actin-like protein more closely resembles Act2 than actin (48 versus 42% identity over 150 residues) but, more significantly, has a similar insertion at residue 42. Additionally, both proteins are most divergent from actin in the same regions (residues 39–49 and 91–103).

The recent resolution of the tridimensional structure of the actin:DNase I complex⁹ and the building of an atomic model of the actin filament¹⁰ allowed us to superimpose the Act2 polypeptide chain onto the structure of actin. By looking for only those residues in the Act2 protein which are different from the consensus actin positions, it is obvious that most of the divergence occurred in the protein loops, leaving the core of the molecule largely unaffected. Hence, we can assume that the overall structure of Act2 protein folds in a very similar way to the actin molecule. Our proposition is greatly strengthened by the fact that, in the core region in the cleft between the two domains, 14 out of the 17 residues involved in ATP and Ca^{2+} binding are strictly conserved (the three remaining ones being substituted by an equivalent amino acid). The Act2 protein must therefore bind and probably hydrolyse ATP as the actins do.

But because the amino-acid insertions as well as other significant changes reside exactly at the interface between actin monomers^{9,10}, we hypothesize that the Act2 protein must have some particular property. Is the Act2 protein unable to polymerize into filaments or does it promote their disassembly? The estimated stoichiometry between the Act2 protein and actin is consistent with such a regulatory role. Alternatively, the Act2 protein could form structures different from the classic actin filaments or cables. Is it associated with the ring of 10-nm filaments¹⁸ located at the mother-bud junction as the phenotype of null mutants is consistent with a role for the Act2 protein in cytokinesis? Analysis of *act2* mutants and cellular localization of the Act2 protein should help to answer these questions. The DNase I and myosin-binding regions are drastically altered and the ubiquitous histidine 73, which is almost universally methylated¹⁹ in actins, is substituted by an asparagine in both Act2 and the nematode actin-like protein. By contrast, arginine 177, which is ADP-ribosylated²⁰ by clostridial toxins, as well as residues 117 (glutamic acid), 119 (methionine) and 355 (methionine) involved in the binding of phallotoxins²¹, are all present in the Act2 protein.

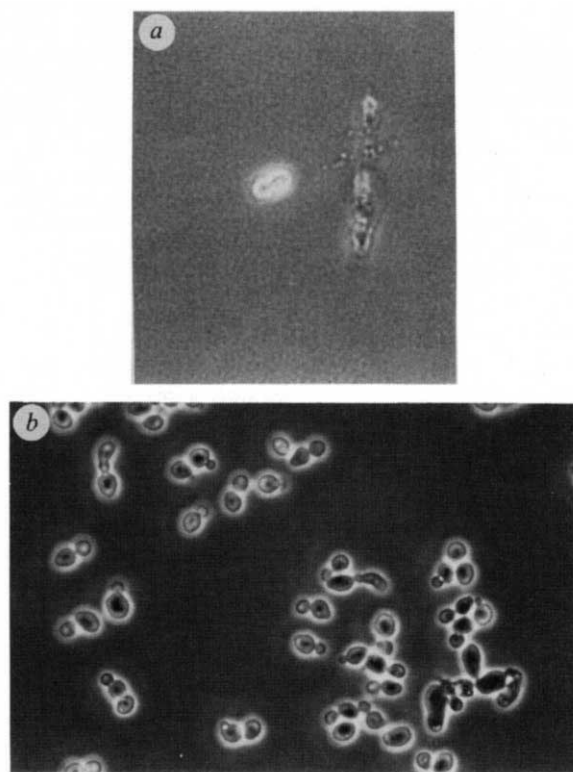


FIG. 2 Microscope analysis of *ACT2*-disrupted yeast cells. *a*, Arrest phenotype of *act2::URA3* null mutants. After disruption of the *ACT2* gene and sporulation of the yeast cells (at 25 °C), tetrads were dissected using a micromanipulator and nongrowing *act2::URA3* null haploids were observed under a phase-contrast microscope (magnification $\times 400$) after 6 days of incubation at 25 °C of the dissection plate. All *act2* null mutants arrested with a large bud before completion of the first cell-division cycle indicating that the *ACT2* gene encodes an essential protein required at a specific stage late in the cell cycle. *b*, Phenotype of *ACT2/act2::URA3* diploid cells. Diploid cells with one copy of their *ACT2* gene disrupted by insertion of the *URA3* marker were grown overnight in minimal medium, transferred to rich medium for 6 h, sonicated for 20 s at low cell density and then observed under a phase-contrast microscope (magnification $\times 400$). About 20% of the cells are enlarged and carry multiple buds separated from the mother cell by a rather long neck. Continuity of the cell wall/membrane between the mother and daughter cells could be assessed by moving the focal plane of the microscope and was confirmed by micromanipulation. This multibudded phenotype is reminiscent of an altered cytoskeleton⁸ and/or a defect in cytokinesis¹⁸.

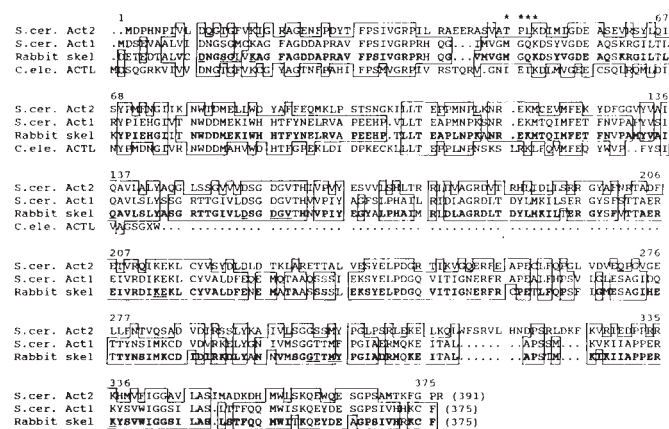


FIG. 3 Alignment of the Act2 protein sequence with yeast cytoplasmic actin^{3,4} (S. cer. Act1), rabbit skeletal α -actin¹⁷ (Rabbit skel.) and *C. elegans* actin-like protein (C. ele. ACTL; partial sequence, Genbank/EMBL accession number M75768). Compared with actins, the putative Act2 protein has an 11 amino-acid insertion at residue 320; four other smaller insertions occur at residues 42, 102, 348 and 375, whereas the N terminus is one residue shorter. To help the comparison with the three-dimensional structure⁹, numbering refers to the sequence of rabbit α -actin. Residues identical in two or more contiguous sequences are boxed. The putative Cdc28/p34^{cdc2} phosphorylation site²² of the Act2 protein is indicated by asterisks. Underlined amino acids are those involved in ATP and Ca²⁺ binding⁹. Consensus residues (that is, those that are present in 12 out of 15 prealigned actin sequences *S. cerevisiae* ACT1 and ACT2, *S. pombe*, *Physarum*, *Thermomyces*, *Tetrahymena*, *Oxytricha*, *Trypanosoma*, soybean, maize, *Drosophila*, sea urchin, *Xenopus*, mouse and human) are shown in bold. Except for the yeast ACT2 and nematode actin-like, all sequences were obtained from the Genbank/EMBL/PIR databases. The alignment and consensus were created with the Pileup and Pretty subroutines of the UWGCG package²⁸.

The most striking motif found in the Act2 polypeptide and in none of the classic actins is certainly a consensus sequence (TPLK; single-letter amino-acid code) for phosphorylation by the mitotic Cdc28/cyclin protein kinase complex²² (p34^{cdc2} in other organisms). This motif (residues 49–52) is located in an exposed loop that, in the atomic model of the actin filament, interacts with the above monomer along the two-start actin helix^{9,10}. The recent isolation of G2 cyclins from *S. cerevisiae*^{23,24} opens perspective for a role of the Cdc28 protein kinase late in the cell cycle. As our gene disruption experiments suggest a cell-cycle regulated function for the Act2 protein (that is, cytokinesis), a very attractive working hypothesis is that the putative phosphorylation of the Act2 protein by the Cdc28 protein kinase could bring about the cytoskeletal reorganization taking place during mitosis. □

Received 22 August; accepted 18 October 1991.

- Pollard, T. D. & Cooper, J. A. *Rev. Biochem.* **55**, 987–1035 (1986).
- Drubin, D. G. *Cell Motil. Cytoskel.* **15**, 7–11 (1990).
- Gallwitz, D. & Sures, I. *Proc. natn. Acad. Sci. U.S.A.* **77**, 2546–2550 (1980).
- Ng, R. & Abelson, J. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3912–3916 (1980).
- Shortle, D., Haber, J. E. & Botstein, D. *Science* **217**, 371–373 (1982).
- Kilmartin, J. V. & Adams, A. E. M. *J. Cell Biol.* **98**, 922–933 (1984).
- Adams, A. E. M. & Pringle, J. R. *J. Cell Biol.* **98**, 934–945 (1984).
- Novick, P. & Botstein, D. *Cell* **40**, 405–416 (1985).
- Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. *Nature* **347**, 37–44 (1990).
- Holmes, K. C., Popp, D., Gebhard, W. & Kabsch, W. *Nature* **347**, 44–48 (1990).
- Snyder, M., Elledge, S., Sweetser, D., Young, R. A. & Davis, R. W. *Meth. Enzym.* **159**, 107–128 (1987).
- Legrain, P. *et al. Molec. Gen. Genet.* **225**, 199–202 (1991).
- Mortimer, R. K. & Schild, D. *Microbiol. Rev.* **49**, 181–212 (1985).
- Greer, C. & Schekman, R. *Molec. cell. Biol.* **2**, 1270–1278 (1982).
- Huffaker, T. C., Hoyt, M. A. & Botstein, D. *A. Rev. Genet.* **21**, 259–284 (1987).
- Adams, M. D. *et al. Science* **252**, 1651–1656 (1991).
- Elzinga, M., Collins, J. H., Kuehl, W. M. & Adelstein, R. S. *Proc. natn. Acad. Sci. U.S.A.* **70**, 2687–2691 (1973).
- Haarer, B. K. & Pringle, J. R. *Molec. cell. Biol.* **7**, 3678–3687 (1987).
- Solomon, L. R. & Rubenstein, P. A. *J. biol. Chem.* **262**, 11382–11388 (1988).
- Aktories, K. & Wegner, A. *J. Cell Biol.* **109**, 1385–1387 (1989).
- Vandekerckhove, J., Deboen, A., Nassal, M. & Wieland, T. *EMBO J.* **4**, 2815–2818 (1985).

- Moreno, S. & Nurse, P. *Cell* **61**, 549–551 (1990).
- Surana, U. *et al. Cell* **65**, 145–161 (1991).
- Ghiara, J. B. *et al. Cell* **65**, 163–174 (1991).
- Langford, C. J. & Gallwitz, D. *Cell* **33**, 519–527 (1983).
- Chasman, D. I. *et al. Genes Dev.* **4**, 503–514 (1990).
- Dale, R. M. K., McClure, B. A. & Houchins, J. P. *Plasmid* **13**, 31–40 (1985).
- Devereux, J., Haeblerli, P. & Smithies, P. *Nucleic Acids Res.* **12**, 387–395 (1984).
- Rothstein, R. J. *Meth. Enzym.* **101**, 202–211 (1983).
- Sherman, F., Fink, G. R. & Hicks, J. B. *Methods in Yeast Genetics* (Cold Spring Harbor Laboratory, New York, 1986).
- McKnight, S. L. & Kingsbury, R. *Science* **217**, 316–324 (1982).

ACKNOWLEDGEMENTS. We thank F. Lacroute (CGM, Gif-sur-Yvette) for yeast strains and plasmids and F. Exinger (IBMC, Strasbourg) for help with the yeast dissection techniques. M. L. Bach, X. Wilhelm (IBMC), G. Burger (Orsay), F. Lang (Montreal), T. Dinger (Erlangen) for the gift of various genomic DNAs, M. Snyder (New Haven) for the yeast genomic library, B. Guard (Gif-sur-Yvette) for the chromosome blot, O. Poch and F. deRaipont (IBMC) for the cloning of the *Drall-Sphl* subfragment of the ACT2 gene, B. Winsor for useful comments and critical reading of the manuscript and G. Dirheimer for encouragement. This work was supported by the CNRS. The nucleotide sequence data reported will appear in the EMBL, Gen Bank and DDBJ Nucleotide Sequence databases under the accession number X61502.

Polar and lateral flagellar motors of marine *Vibrio* are driven by different ion-motive forces

Tatsuo Atsumi*, Linda McCarter† & Yasuo Imae*‡

* Department of Molecular Biology, Faculty of Science, Nagoya University, Chikusa-ku, Nagoya 464-01, Japan

† The Agouron Institute, 505 Coast Boulevard South, La Jolla, California 92037, USA

VARIOUS species of marine *Vibrio* produce two distinct types of flagella, each adapted for a different type of motility¹. A single, sheathed polar flagellum is suited for swimming in liquid medium, and numerous unsheathed lateral flagella, which are produced only under viscous conditions, are suited for swarming over viscous surfaces^{2,3}. Both types of flagella are driven by reversible motors embedded in the cytoplasmic membrane. Here we report that the energy source for the polar flagellar motor of *Vibrio parahaemolyticus* is the sodium-motive force, whereas the lateral flagellar motors are driven by the proton-motive force. This is evidence that two distinct types of flagella powered by different energy sources are functionally active in one cell.

To investigate the energy sources for the two types of flagellar motors, we used mutant strains of *V. parahaemolyticus* having only one type of flagella: swimmer strain RS313 (Fla⁺Laf⁻) bearing a single polar flagellum only and swarmer strain ML34 (Fla⁻Laf⁺) bearing lateral flagella only³. The lateral flagella in ML34 are constitutively produced even under non-viscous conditions³. As the polar flagellar motor of the related bacterium *Vibrio alginolyticus* is known to be powered by the sodium-motive force^{4,5}, we first examined the Na⁺ dependence of motility of *V. parahaemolyticus* mutant strains. As shown in Fig. 1a, RS313 with only a polar flagellum showed clear Na⁺-dependent motility, suggesting that the polar flagellar motor of *V. parahaemolyticus* is also driven by the sodium-motive force. In contrast, the motility of ML34 was not affected by the presence or absence of Na⁺ in the medium (Fig. 1a). These results suggest that the energy source for the lateral flagellar motors is not the sodium-motive force, although these motors may possibly require only a very low concentration of Na⁺, which may be present as a contaminant from various sources during experiments. To eliminate this possibility, we tested the amiloride sensitivity of motility in both strains, as amiloride is known to inhibit various Na⁺-driven flagellar motors, including the polar flagellar motor of *V. alginolyticus*, in a competitive manner with Na⁺ in the medium^{6–8}. As shown in Fig. 1b, 2.5 mM amiloride completely inhibited the motility of RS313 but not of ML34.

‡ To whom correspondence should be addressed.

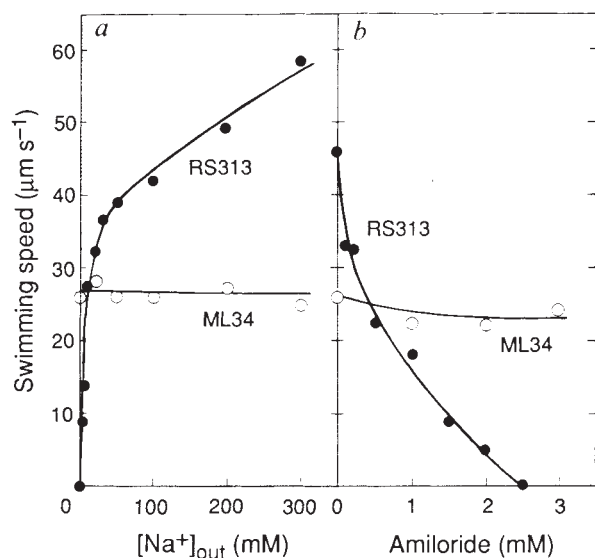


FIG. 1 Na^+ requirement and amiloride sensitivity of the polar and lateral flagellar motors in *V. parahaemolyticus*. The strains used were RS313 (*laf-313::lux tet'*) and ML34 (*fla-340::lac tet'*). Cells were grown in 2216 medium (Difco) at 30 °C with (for ML34) or without (for RS313) 10% polyvinylpyrrolidone-360 (PVP), which increases the viscosity of the medium³. Swimming speed of the cells was measured at 25 °C as described previously¹³. For ML34, 10% polyvinylpyrrolidone (PVP) was supplemented in all media to minimize the detachment of lateral flagella, although essentially the same results were obtained in media without PVP. *a*, Na^+ dependence of motility. Cells were washed with medium A (50 mM HEPES-choline buffer (pH 7.0), 10 mM MgCl_2 , 5 mM glucose) supplemented with 300 mM KCl, and then diluted 100-fold with the same medium supplemented with 300 mM salts consisting of various concentrations of NaCl and KCl. Motility was measured within 1 min after dilution. *b*, Amiloride sensitivity of motility. Cells were washed and resuspended with medium A supplemented with 50 mM NaCl and 250 mM KCl. Motility was measured within 1 min after the addition of various concentrations of amiloride.

Thus, the results again suggested a difference in the coupling ions for the polar and lateral flagellar motors.

A wild-type strain of *V. parahaemolyticus*, BB22, grown under viscous conditions, such as in the presence of 10% polyvinylpyrrolidone, has many lateral flagella in addition to a polar flagellum^{2,3}. Such cells showed good motility without the addition of Na^+ but their swimming speed was significantly increased with increasing concentrations of Na^+ in the medium (Fig. 2). In contrast, BB22 cells grown in the absence of polyvinyl-

TABLE 1 Effect of CCCP and HQNO on the swimming speed of *V. parahaemolyticus* RS313 and ML34

pH of medium*	CCCP (μM)	HQNO (μM)	Swimming speed ($\mu\text{m s}^{-1}$)	
			RS313 (Fla^+Laf^-)	ML34 (Fla^-Laf^+)
7.0	0	0	60	26
	20	0	7	0
8.5	0	0	58	23
	20	0	51	0
	0	20	50	25
	20	20	0	0

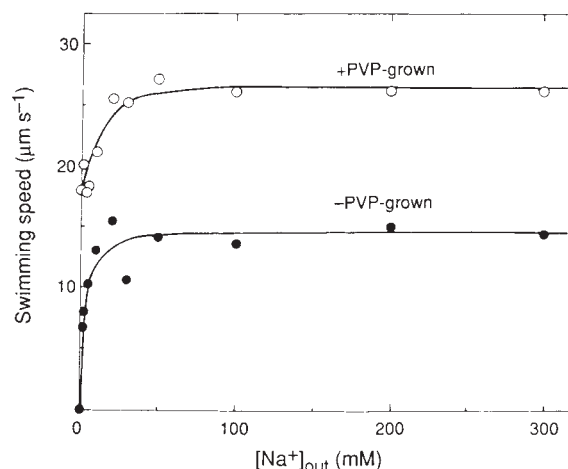
* Motility was measured in 2216 medium, and the pH of the medium was adjusted by adding either HCl to pH 7.0 or K_2CO_3 to pH 8.5.

pyrrolidone have only a polar flagellum, and showed a clear Na^+ -dependent motility in a similar fashion to RS313 (Fig. 2). Thus, the difference in the coupling ions for the two types of motors is also observed for the wild-type strain, and under viscous conditions both the polar and lateral flagellar motors seemed to function to drive the cell.

We then examined whether the lateral flagellar motors are powered by the proton-motive force, as is the case for the motors of bacteria such as *Escherichia coli* and *Bacillus subtilis*^{9,10}. *V. parahaemolyticus* is peculiar in having, in addition to a respiration-coupled H^+ pump, a respiration-coupled Na^+ pump, which is active in the alkaline range^{11,12}. As shown in Table 1, when carbonylcyanide *m*-chlorophenylhydrazone (CCCP), a protonophore, was added to RS313 with a polar flagellum, motility was lost at pH 7 but not affected at pH 8.5. Even at pH 8.5, CCCP diminishes the proton-motive force but the Na^+ pump, which is active at this pH, generates the CCCP-insensitive sodium-motive force¹¹. Consistent with this, further addition of 2-heptyl-4-hydroxyquinoline-*N*-oxide (HQNO), a specific inhibitor of the Na^+ pump¹¹, caused complete loss of motility of RS313. These results confirm that the polar flagellar motor is powered by the sodium-motive force. In the case of laterally flagellated ML34, motility was completely inhibited by 20 μM CCCP at both pH 8.5 and pH 7. These results indicate that the lateral flagellar motor is not dependent on the sodium-motive force but is driven by the proton-motive force.

The behaviour of wild-type strain BB22 grown in non-viscous media was similar to RS313: motility was amiloride-sensitive but CCCP-resistant at pH 8.5. As shown in Fig. 3, however, BB22 grown in viscous media behaved differently. The swimming speed at pH 8.5 was only partly reduced by the addition of either CCCP or amiloride; however, it was dramatically

FIG. 2 Na^+ dependence of motility of *V. parahaemolyticus* wild-type strain (BB22) grown in 2216 medium supplemented with or without 10% PVP. Experiments were carried out as in Fig. 1*a* by using the motility medium supplemented with 10% PVP.



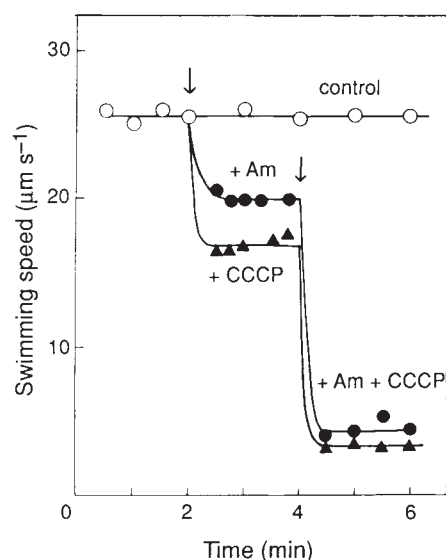


FIG. 3 Effect of amiloride and CCCP on the motility of wild-type cells (BB22) grown in viscous medium (2216 medium supplemented with 10% PVP). Motility was measured as in Fig. 1*b*, except that medium B (25 mM Tris-HCl buffer (pH 8.5), 10 mM $MgCl_2$, 5 mM glucose) supplemented with 50 mM NaCl, 250 mM KCl and 10% PVP was used for the experiments. At the point indicated by first arrow, 3 mM amiloride (+Am) or 50 μ M CCCP (+CCCP) was added, and at the second arrow, the other drug was added.

reduced by the addition of both together. These results indicate that in the dually flagellated BB22 cell, the polar flagellum is powered by the sodium-motive force and the lateral flagella are powered by the proton-motive force, and both function additively. When the wild-type strain 138-2 of *V. alginolyticus* was grown in the presence of 10% PVP, the pattern of motility was similar to that of BB22 shown in Figs 2 and 3, indicating that the two types of flagellar motors of *V. alginolyticus* are also powered by the different ion-motive forces.

The swimming speed of RS313 or BB22 with only a polar flagellum was about $60 \mu\text{m s}^{-1}$ in liquid media but was reduced to $15 \mu\text{m s}^{-1}$ in viscous media (see Fig. 2). In contrast, the swimming speed of laterally flagellated ML34 was not affected by the viscosity of the medium: speed remained constant (about $25 \mu\text{m s}^{-1}$). Thus, the polar flagella are adapted for swimming in liquid media, whereas lateral flagella seem suited for movement in viscous environments. These results are consistent with the lifestyle of the organism. Marine *Vibrio* can be isolated from a variety of habitats, freely swimming or attached to animate and inanimate surfaces¹.

How has the organism come to use two different energy sources to power movement under different circumstances? An evolutionary explanation could be that a marine bacterium with a motility system adapted for swimming in sea water might have acquired a second motility system from a bacterium such as *E. coli* which has the H^+ -driven flagellar system. The structural differences observed between polar and lateral flagella could support this idea. A functional explanation could be that the pH of animate or inanimate surfaces is more acidic than that of sea water (pH ~ 8.2). The respiration-coupled Na^+ -pump activity would then be reduced, and the use of H^+ -driven motors for swarming over surfaces would have merit in terms of cellular energetics. The polar flagellum is also thought to function as a dynamometer, controlling the swarmer cell differentiation by sensing forces that interfere with its own rotation^{2,3}. The change in the coupling ion for the motors may also have some relation to this dynamometer function. □

Received 21 August; accepted 17 October 1991.

- Allen, R. D. & Baumann, P. *J. Bact.* **107**, 295–302 (1971).
- McCarter, L. & Silverman, M. *Molec. Microbiol.* **4**, 1057–1062 (1990).
- McCarter, L., Hilmen, M. & Silverman, M. *Cell* **54**, 345–351 (1988).
- Chernyak, B. V., Dibrov, P. A., Glagolev, A. N., Sherman, M. Y. & Skulachev, V. P. *FEBS Lett.* **164**, 38–42 (1983).
- Tokuda, H. et al. *J. Biochem.* **103**, 650–655 (1988).
- Sugiyama, S., Cragoe, E. J. Jr & Imae, Y. *J. Biol. Chem.* **263**, 8215–8219 (1988).
- Atsumi, T., Sugiyama, S., Cragoe, E. J. Jr & Imae, Y. *J. Bact.* **172**, 1634–1639 (1990).
- Imae, Y. & Atsumi, T. *J. bioenerg. Biomembr.* **21**, 705–716 (1989).
- Manson, M. D., Tedesco, P., Berg, H. C., Harold, F. M. & Van der Drift, C. *Proc. natn. Acad. Sci. U.S.A.* **74**, 3060–3064 (1977).
- Matsuura, S., Shioi, J. & Imae, Y. *FEBS Lett.* **82**, 187–190 (1977).
- Tokuda, H. & Umemoto, T. *J. Biol. Chem.* **257**, 10007–10014 (1982).
- Tsuchiya, T. & Shinoda, S. *J. Bact.* **162**, 794–798 (1985).
- Yoshida, S., Sugiyama, S., Hojo, Y., Tokuda, H. & Imae, Y. *J. Biol. Chem.* **265**, 20346–20350 (1990).

ACKNOWLEDGEMENTS. We thank M. Silverman and I. Kawagishi for suggestions and review of this manuscript, and N. Nishio for drawing. This work was supported by a Grant-in-Aid from the Ministry of Education, Science and Culture, Japan (Y.I.) and the NIH (L.M.).

Three-dimensional heteronuclear NMR studies of RNA

Edward P. Nikonowicz & Arthur Pardi

Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309-0215, USA

MULTIDIMENSIONAL heteronuclear NMR has revolutionized solution structure determinations of proteins^{1–3}. But this technique has not been applied to nucleic acids because of difficulties in the synthesis of isotopically (¹³C and/or ¹⁵N) labelled molecules. Here we report the application of three-dimensional heteronuclear NMR to the study of a uniformly ¹³C/¹⁵N or ¹⁵N-labelled RNA duplex of defined sequence. These experiments simplify resonance assignment and the analysis of proton–proton nuclear Overhauser effects⁴ (and therefore distance information) in the molecule. Our results show that it is now possible to determine the structures of larger and more complex RNAs using multidimensional heteronuclear NMR.

The discovery that RNA can catalyse chemical reactions has spawned many investigations of the functions of RNA in living systems^{5–7}. Although X-ray crystal and NMR solution structures of hundreds of biologically active enzymes and proteins are known⁸, transfer RNA is still the only sizeable RNA for which a high-resolution three-dimensional structure has been determined. NMR spectroscopy is a powerful technique for the determination of protein structures in solution, but has been less useful for nucleic acids because of the limited spectral resolution of the proton resonances in DNA and RNA^{4,9,10}. For example, excluding the H1' resonances, all the sugar protons in RNAs resonate in a very narrow (~ 1.5 p.p.m.) spectral region. Fortunately, the resolution of the NMR experiment can be dramatically improved by spreading the spectral data into an increased number of dimensions^{1–3,11}. Figure 1 shows the H1'–H2'–C1' region of a three-dimensional (3-D) HCCH-COSY (refs 12, 13) NMR experiment of a 99% uniformly ¹³C/¹⁵N-labelled RNA duplex, 5'-(pGGCGCUUGCGUC)₂-3'. This experiment provides through-bond correlations, and therefore these cross-peaks identify H1' and H2' resonances belonging to the same sugar. The very small (~ 1 Hz) H1'–H2' three-bond coupling normally found in an A-form RNA helix often makes it impossible to observe H1'–H2' correlations in two-dimensional (2-D) ¹H spectra of RNAs^{4,9,10}. However, uniform incorporation of ¹³C allows very efficient transfer of magnetization between protons on neighbouring carbons^{12–14} through the large one-bond ¹H–¹³C (145–160 Hz) coupling and ¹³C–¹³C (38–45 Hz) coupling. Additionally, a large increase in the resolution of the ¹H–¹H cross peaks is obtained by extending these data into a third, ¹³C, frequency dimension, thereby greatly simplifying the analysis of the spectrum. Therefore isotope labelling of RNA allows

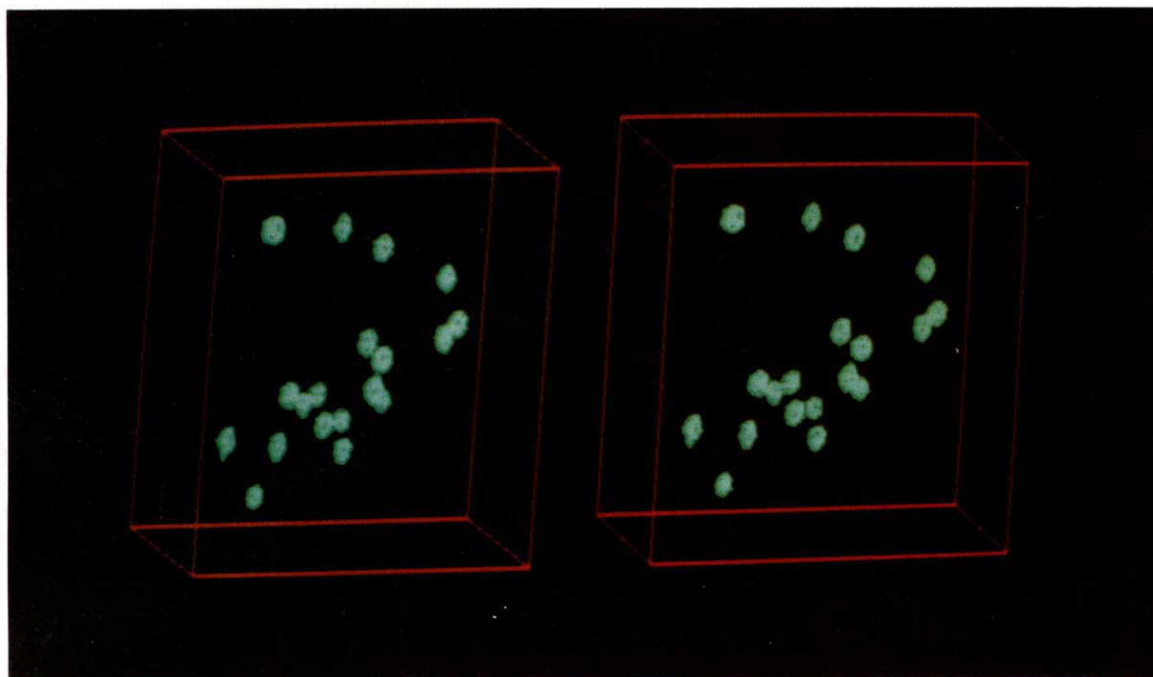


FIG. 1 Stereo view of the H1'-H2'-C1' region of the 3-D HCCH-COSY spectrum of the uniformly $^{13}\text{C}/^{15}\text{N}$ -labelled RNA duplex $\text{r}(\text{pGGCGCUUGCUC})_2$ at 1.7 mM single strands in 99.996% $^2\text{H}_2\text{O}$, 0.14 M NaCl, 10 mM potassium phosphate, 0.1 mM EDTA, pH 6.8 at 30°C . This particular sequence was chosen because it has the potential to form a hairpin structure with a commonly occurring CUUG tetranucleotide loop motif²¹. Under the conditions used here only the duplex is observed in solution. Because of the small spectral width of the 3-D spectrum, the cytidine and uridine H5-H6 cross-peaks have been folded into this region.

METHODS. The 99% $^{13}\text{C}/^{15}\text{N}$ or ^{15}N labelled samples were prepared by *in*

vitro transcription using T7 RNA polymerase²². Isotopically labelled rNTPs were prepared by extracting the ribosomal RNA from *Escherichia coli* grown on 99% $^{13}\text{C}/^{15}\text{N}$ - or ^{15}N -enriched medium, degrading the rRNA to 5' rNMPs with nuclease P1 and then synthesizing the NTPs using enzymatic procedures²³. The 3-D HCCH-COSY experiment was acquired on a Varian UNITY-500 NMR spectrometer as described^{12,13} using sweep widths of 1,600, 2,800 and 3,200 Hz over 56, 28 and 1,024 complex points in the t_1 (^1H), t_2 (^{13}C) and t_3 (^1H) frequency dimensions, respectively. All the 3-D data were transferred to a Silicon Graphics 4D/25GT computer and processed with the FELIX program (Hare Research).

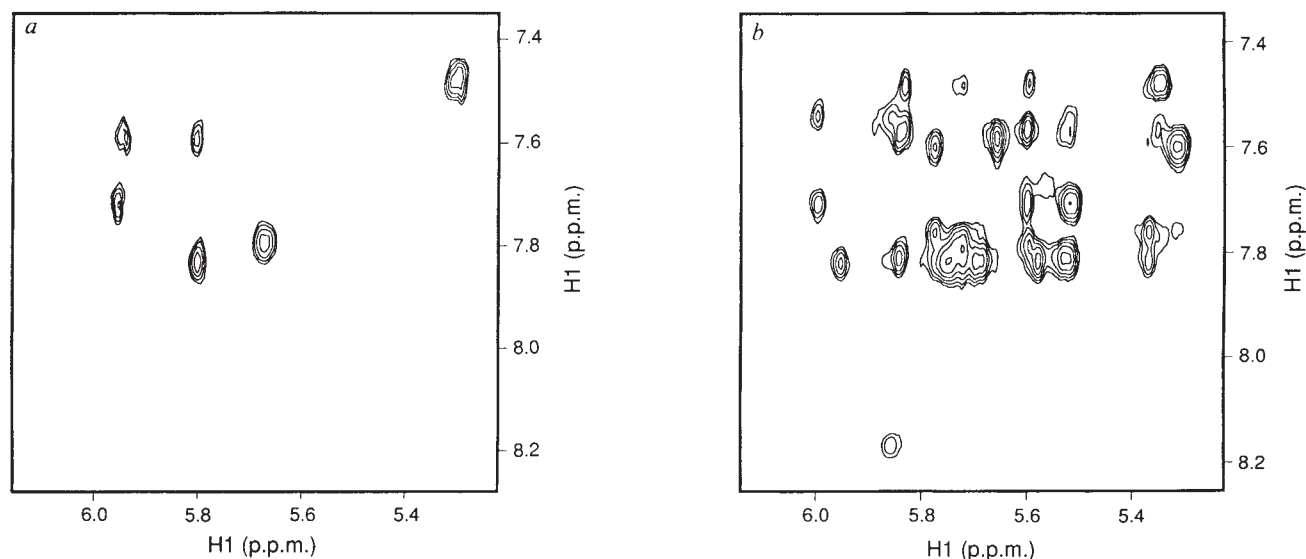
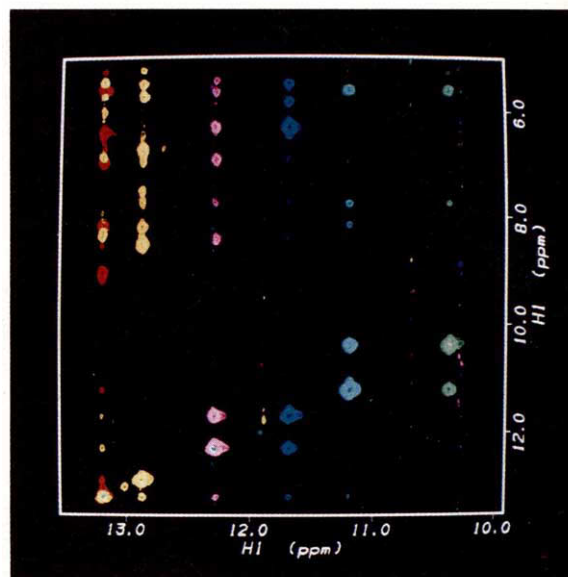


FIG. 2 *a*, Contour plot of the H1' to aromatic proton region of a 2-D slice (along the ^{13}C dimension) of the 3-D ^{13}C NOESY-HMQC spectrum of the RNA duplex using the sample conditions described in the legend to Fig. 1. *b*, Contour plot of the same region of a 2-D NOESY spectrum of the unlabelled RNA duplex using the buffer conditions described in the legend to Fig. 1, except at 0.30 M NaCl. Some crosspeaks have minor chemical shift changes due to slightly different salt conditions for the labelled and unlabelled samples. Comparison of these spectra shows that data from the 3-D

spectrum is much simplified compared with the 2-D spectrum.

METHODS. The 3-D experiment was acquired with a previously described pulse sequence¹⁶, using sweep widths of 3,200, 5,300 and 3,200 Hz over 56, 61 and 512 complex points in the t_1 (^1H), t_2 (^{13}C) and t_3 (^1H) frequency dimensions, respectively, with a 400-ms mixing time. The 2-D experiment was acquired with similar spectral parameters, except that additional points were collected in both frequency dimensions.

FIG. 3 Superposition of contour plots of several 2-D slices of the 3-D NOESY-HMQC spectrum of the uniformly ^{15}N -labelled RNA duplex. Sample conditions were the same as those for Fig. 1, except that the spectrum was run in 90% $\text{H}_2\text{O}/10\%$ $^2\text{H}_2\text{O}$ at 10°C . The 2-D slices at 145.5, 144.7/144.5, 152.2, 142.0, 154.6 and 156.2 p.p.m. in the ^{15}N dimension are plotted in red, yellow, green, blue, cyan and pink, respectively. The strong imino-to-imino proton crosspeaks arise from the G-U and U-U base pairs (see text). METHODS. The 3-D spectrum was acquired with a modified NOESY-HMQC pulse sequence¹⁶ that contained a jump and return water suppression pulse²⁴ to reduce the large water signal. The spectrum used sweep widths of 10,000, 1,200 and 10,000 Hz over 100, 13 and 1,024 complex points in the t_1 (^1H), t_2 (^{15}N) and t_3 (^1H) frequency dimensions, respectively, with a 350-ms mixing time.



spectral simplification both through 3-D (or 4-D) heteronuclear NMR experiments^{1-3,11} and through the efficient observation of ^1H - ^1H scalar correlations¹²⁻¹⁴.

The sequential resonance assignment procedure in nucleic acids and the subsequent structure determination process rely on observation of proton-proton nuclear Overhauser effects (NOEs) that yield through-space distance information on protons separated by less than $5.0 \text{ \AA}$ (refs 4, 9, 10). Figure 2 shows the $\text{H1}'$ to aromatic proton region of a single 2-D slice of a 3-D ^{13}C NOESY-HMQC spectrum^{15,16} of the RNA duplex. This 3-D experiment spreads the proton-proton NOE data into the third ^{13}C dimension. Also shown is the standard 2-D NOESY spectrum illustrating the huge increase in resolution inherent in the 3-D experiment. This 3-D experiment allows straightforward assignment and NOE analysis of this RNA^{2,3}, and also yields more distance information owing to reduction of crosspeak overlap in the spectrum (E.N. and A.P., unpublished results).

Our 3-D $^1\text{H}/^{13}\text{C}$ experiments facilitate resonance assignment and local structure determination of RNAs. But as most active RNAs have a tertiary structure¹⁷, it is important to identify these tertiary interactions. As these interactions usually involve base-base hydrogen-bonding of a nitrogen-bound hydrogen¹⁷, a 3-D ^{15}N NOESY-HMQC experiment can provide valuable information for RNA tertiary structure. Figure 3 shows a plot of the imino-proton to imino/aromatic-proton region of a 3-D NOESY-HMQC spectrum^{15,16} for the RNA duplex in 90% H_2O . A series of 2-D slices along different ^{15}N frequencies are plotted in different colours. The two imino protons at 13.20 p.p.m. have identical ^1H chemical shifts and therefore are unresolved in a

^1H - ^1H 2-D NOESY spectrum. But as their imino nitrogens resonate at different frequencies, their NOEs are readily separated in the 3-D experiment, making it possible to extract structural data that were unobtainable from the 2-D spectrum. An interesting feature of this spectrum is its unusual pattern of imino-proton crosspeaks. The large crosspeaks involving imino protons at 12.28/11.65 and 11.20/10.37 p.p.m. indicate the presence of a G-U wobble base pair and a U-U base pair respectively^{9,17}. The presence of the U-U base pair is surprising, but confirms thermodynamic data indicating that U-U base pairs stabilize duplex formation in RNAs²⁵. Our NMR results show that 3-D experiments on ^{15}N -labelled RNAs simplify the identification of standard Watson-Crick base pairs and also help to identify unusual or potential tertiary base-base interactions in RNAs.

In summary, we report a novel method for simplifying the resonance assignment and structural analysis of RNAs. Although we used a relatively short RNA fragment, the synthesis of isotopically labelled RNAs of defined sequence and the multidimensional NMR experiments are generally applicable. These techniques will be even more valuable for studies of larger RNAs, especially those beyond the capabilities of standard NMR techniques. We are at present applying multidimensional heteronuclear NMR to the structure of the hammerhead catalytic RNA domain (relative molecular mass 15,000)^{18,19} and are using yeast tRNA^{Phe} (relative molecular mass 25,000)²⁰ as a test system for optimizing the technique to characterize RNA tertiary structure. Our approach should significantly increase the structural information available on biologically active RNAs. □

Received 29 July; accepted 17 September 1991.

1. Fesik, S. W. & Zuiderweg, E. R. P. *Q. Rev. Biophys.* **23**, 97-131 (1990).
2. Marion, D. et al. *Biochemistry* **28**, 6150-6156 (1989).
3. Clore, G. M. & Gronenborn, A. M. *Science* **252**, 1390-1399 (1991).
4. Wüthrich, K. *NMR of Proteins and Nucleic Acids* (Wiley, New York, 1986).
5. Cech, T. C., Zaug, A. J. & Grabowski, P. J. *Cell* **27**, 487-496 (1981).
6. Cech, T. C. *Science* **236**, 1532-1539 (1987).
7. Cech, T. C. *Molecular Biology of RNA* (Liss, New York, 1989).
8. Fersht, A. *Enzyme Structure and Mechanism* 389-442 (Freeman, New York, 1985).
9. Van de Ven, F. J. M. & Hilbers, C. W. *J. Biochem.* **178**, 1-38 (1988).
10. Patel, D. J., Shapiro, L. & Hare, D. *Q. Rev. Biophys.* **20**, 35-112 (1987).
11. Clore, G. M., Kay, L. E., Bax, A. & Gronenborn, A. M. *Biochemistry* **30**, 12-18 (1991).
12. Kay, L. E., Ikura, M. & Bax, A. *J. Am. Chem. Soc.* **112**, 888-889 (1990).
13. Clore, G. M., Bax, A., Driscoll, P. C., Wingfield, P. T. & Gronenborn, A. M. *Biochemistry* **29**, 8172-8184 (1990).
14. Fesik, S. W., Eaton, H. L., Olejniczak, E. T. & Zuiderweg, E. R. P. *J. Am. Chem. Soc.* **112**, 886-888 (1990).
15. Mueller, L. *J. Am. Chem. Soc.* **101**, 4481-4484 (1979).

16. Bax, A. et al. *J. magn. Res.* **86**, 304-318 (1990).
17. Saenger, W. *Principles of Nucleic Acid Structure* (Springer, New York, 1984).
18. Uhlenbeck, O. C. *Nature* **328**, 596-600 (1987).
19. Heus, H. A. & Pardi, A. *J. molec. Biol.* **217**, 113-124 (1991).
20. Hall, K., Sampson, J., Uhlenbeck, O. C. & Redfield, A. *Biochemistry* **28**, 5794-5801 (1989).
21. Woese, C. R., Winker, S. & Gutell, R. R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 8467-8471 (1990).
22. Milligan, J. F., Groebe, D. R., Witherell, G. W. & Uhlenbeck, O. C. *Nucleic Acids Res.* **15**, 8783-8798 (1987).
23. Simon, E. S., Grabowski, S. & Whitesides, G. M. *J. org. Chem.* **55**, 1834-1841 (1990).
24. Plateau, P. & Gueron, M. *J. Am. Chem. Soc.* **104**, 7310-7311 (1982).
25. Santa Lucia, J., Kierzek, R. & Turner, D. H. *Biochemistry* **30**, 8242-8251 (1991).

ACKNOWLEDGEMENTS. We thank L. Mueller (Bristol-Meyers Squibb) for advice in optimizing the 3-D NMR experiments, K. Hall for access to the 500 MHz NMR spectrometer at Washington University, L. Mueller, M. Goriach (University of Pennsylvania) and M. Betlach (Synergen) for isotopically labelled cells, and A. Sirr and L. Baer for assistance in preparation of the isotopically labelled NTPs. This work was supported in part by the NIH, and we also thank the W. M. Keck Foundation for their generous support of RNA science on the Boulder campus.